Available online at www.sciencedirect.com
SCIENCE DIRECT®

Image and Vision Computing xx (2006) 1–13

image AND vision COMPUTING
www.elsevier.com/locate/imavis

Attention links sensing to recognition

Albert L. Rothenstein, John K. Tsotsos *

Department of Computer Science, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3

Received 29 July 2004; received in revised form 12 July 2005; accepted 15 August 2005

Abstract

This paper presents arguments that explicit strategies for visual attentional selection are important for cognitive vision systems, and shows that a number of proposals currently exist for exactly how parts of this goal may be accomplished. A comprehensive survey of approaches to computational attention is given. A key characteristic of virtually all the models surveyed here is that they receive significant inspiration from the neurobiology and psychophysics of human and primate vision. This, although not necessarily a key component of mainstream computer vision, seems very appropriate for cognitive vision systems given a definition of the topic that always includes the goal of human-like visual performance. A particular model, the Selective Tuning model, is overviewed in some detail. The growing neurobiological and psychophysical evidence for its biological plausibility is cited highlighting the fact that it has more biological support than other models; it is further claimed that it may form an appropriate starting point for the difficult task of integrating attention into cognitive vision systems.

Keywords: Cognitive vision; Attention; Recognition; Selective tuning

1. Introduction

Regarding cognitive vision, it has been claimed that

> 'The ultimate goal is to develop a general-purpose vision system with the robustness and resilience of the human visual system. In particular, cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events and structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment.' [1].

This paper discusses one small but critical slice of a cognitive computer vision system, that of visual attention. This presentation begins with a brief discussion on a definition for attention followed by an enumeration of the different ways in which attention should play a role in computer vision. A detailed review of current biologically motivated computational models of attention is then presented with the goal of discovering what the key ingredients of attentional processing may be and how the current state-of-the-art deals with them. The Selective Tuning (ST) model [2,3] is then overviewed with an emphasis on its components that are most relevant for cognitive vision, namely the winner-take-all (WTA) processing, the use of distributed saliency and feature binding as a link to recognition.

2. Towards a definition of attention

What is 'attention'? Is there a computational justification for attentive selection? The obvious answer that has been given many times that the brain is not large enough to process all the incoming stimuli, is hardly satisfactory [4]. This answer is not quantitative and provides no constraints on what processing system might be sufficient. Methods from computational complexity theory have formally proved that purely data-directed visual search in its most general form is an intractable problem in any realization [5]. There, it is claimed that visual search is ubiquitous in vision, and thus purely data-directed visual processing is also intractable in general. Those analyses provided important constraints on visual processing mechanisms and led to a specific (not necessarily unique or optimal) solution for visual perception. One of those constraints concerned the importance of attentive processing at all stages

[*] Corresponding author. Tel.: +1 416 736 2100; fax: +1 416 736 5872.
E-mail addresses: albertlr@cs.yorku.ca (A.L. Rothenstein), tsotsos@cs.yorku.ca (J.K. Tsotsos).

0262-8856/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.imavis.2005.08.011

of analysis: the combinatorics of search are too large at each stage of analysis otherwise. Attentive selection based on task knowledge turns out to be a powerful heuristic to limit search and make the overall problem tractable [6]. This conclusion leads to the following view of attention: "attention is a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception". It thus plays a role in all aspects of vision.

Many (the active/animate vision researchers) seem to claim that attention and eye movements are one and the same; certainly none of the biological scientists working on this problem would agree. That one can attend to particular locations in the visual field without eye movements has been known since Helmholtz [7], but eye movements require visual attention to precede them to their goal ([8] surveys relevant experimental work). Both selection goals are needed corresponding to overt and covert attentional fixations described in the perception literature. Active vision, as it has been proposed and used in computer vision, necessarily includes attention as a sub-problem.

3. Attention in computer vision

What is it about attention that makes it one of the easiest topics to neglect in computer vision? Look at the majority of books on computer vision and one will not find attention even mentioned. One surprising example is the book *Active Vision* edited by Aloimonos [9], where one would expect attention to be prominent. The index includes only one entry under the term 'attention'. That entry points to a paper by Sandini et al. [10] that tells us that the task of tracking, or active control of fixation, requires as a first step the detection of the target or focus of attention. How would one go about solving this? In [4], it was shown that with no task knowledge and in a purely data-directed manner, this sub-task of target detection is NP-Complete making it appear as if these authors are attempting to solve a problem that includes known intractable sub-problems. What conclusions can be drawn from such proposals? Is the problem thought to be irrelevant or is it somehow assumed away?

Those who build complete vision application systems invoke attentional mechanisms because they must confront and defeat the computational load in order to achieve the goal of real-time processing (there are many examples, two of them being [11,12]). But the mainstream of computer vision does not give attentive processes, especially task-directed attention, much consideration.

A spectrum of problems requiring attention has appeared [13]: selection of objects, events, tasks relevant for domain, selection of world model, selection of visual field, selection of detailed sub-regions for analysis, selection of spatial and feature dimensions of interest, selection of operating parameters for low level operations. The following is a list of common assumptions that reduce or eliminate the need for attention made in the literature:

- fixed camera systems negate the need for selection of visual field
- pre-segmentation eliminates the need to select a region of interest
- 'clean' backgrounds ameliorate the segmentation problem
- assumptions about relevant features and the ranges of their values reduce their search ranges
- knowledge of task domain negates the need to search a stored set of all domains
- knowledge of which objects appear in scenes negates the need to search a stored set of all objects
- knowledge of which events are of interest negates the need to search a stored set of all events.

The point is that the extent of the search space is seriously reduced before the visual processing takes place, and most often even before the algorithms for solution are designed! However, it is clear that in everyday vision, and certainly in order to understand vision, these assumptions cannot be made. More importantly, the need for attention is broader than simply vision as the above list shows. It touches on the relevant aspects of visual reasoning, recognition, and visual context.

4. Computational models of visual attention

Since the goal is to develop a general-purpose vision system with human-like performance, it is instructive to consider what those human characteristics might be. In the domain of computational solutions for visual attention, this connection has been fruitfully exploited for some time. This section will explore the characteristics of human visual attention performance and describe how modelling efforts to date have attempted to realize that performance, with an emphasis on neural models that attempt to explain the system-level behaviour starting from the response properties of visual neurons.

4.1. Key ingredients

Even a summary review of the relevant literature shows that visual attention is a very broad and fragmented field of study. To account for the wealth of experimental data in a single model is not an easy task, and as such, it is probably not surprising that most models of visual attention focus on explaining particular aspects of this complex phenomenon. To be in a better position to compare the various models and trace their evolution, we will start by listing the issues driving both experimental research and the development of theories and models.

A theory of attention must be able to explain salience, pop out, visual search asymmetries, and a series of other phenomena usually termed 'pre-attentive'. While some authors do not consider these to be components of attention per se, but a mere side effect of pre-processing steps, the mechanisms involved are very likely either shared or at the very least intimately related. Some theories actually claim that the same mechanism is responsible for both salience and attentional selection, usually taking the route of eliminating the centralized representation of saliency (the 'saliency map'), and rely on the competitive dynamics of the system for attentional selection (e.g. [2,14]).

Several forms of top-down information are identified in the literature. Explicit information can take the form of verbal instructions or image cues. Implicit information is based on previously viewed stimuli. The latter can take the form of ‘priming of pop-out’ [15] when subjects respond faster to a feature if recent targets have been the same feature, or ‘contextual cueing’ [16] when subjects learn that the target is more likely to appear at a particular location, even if they are not explicitly aware of this. These and other related issues are discussed in detail by Wolfe [17]. Since experimental evidence shows that high-level information influences attentional deployment, theories of attention must be able to explain the mechanisms of this top-down modulation. There are a number of components to this, from describing the neural mechanisms involved in the modulation and how they interact with the bottom-up processes, to visual object representations and how objects are selected as the subject of attention.

Another important component of a unified theory of attention is attentional selection, i.e. what kinds of stimuli can be attended, and what exactly does it mean that a particular stimulus is selected by attention? These are very broad question, touching on issues such as attending to locations vs. objects vs. features, and the size and shape of the attentional field. Attentional selection is ultimately the interface between perception and consciousness [18], and relatively little is known with any degree of certainty about this interface. Closely related to this is the issue of attentional control, i.e. how are shifts of attention controlled and where do the control signals come from [19]. The reverse of selection needs to be explained, i.e. what happens to stimuli that are not attended? Experimental evidence shows that unattended stimuli can influence behaviour [20], but in general, the processing of unattended stimuli is either incomplete, or incorrectly bound [21–23]. What goes wrong when attention is not present can provide us with very important clues about the role of attention in perception.

After a stimulus has been selected as focus of attention, important questions are how can its reselection be prevented, and under what conditions does reselection occur? As with any other topic in visual attention, opinions are divided, the classical view requiring a mechanism of inhibition of return (IOR) [24] while a controversial study [25] suggests that memory plays only a marginal role in visual search. Most theories and models deal with covert attention, i.e. focusing without eye movements, but in day to day life overt attention, i.e. eye movements that foveate stimuli of interest, dominate, and this adds another layer of complexity to the issue of IOR. Various studies seem to indicate that overt and covert attention are closely related and that the same neuronal structures are involved (e.g. [26]). At least two aspects of this dichotomy need to be addressed: what are the similarities and differences between the two, and in the case of overt shifts of attention, what are the IOR mechanisms that compensate for the saccadic shifts in the retinal images.

Last but not least, what are the implications of attentional selection from an object recognition point of view. Experimental results suggest that attentional load influences our ability to make perceptual judgments differentially, depending on the kinds of stimuli presented [27–29], but the mechanisms at work are far from clear.

All this has to be done accounting for results at all levels, from neurophysiology to behaviour and cognitive science.

4.2. Saliency

Many computational models of visual attention use saliency maps (initially proposed by Koch and Ullman [30]), in some form or another, as a two-dimensional scalar map of values representing the visual saliency of the corresponding location, irrespective of the particular stimulus information that makes the location salient. With this hypothesis, focusing attention to the most salient location is reduced to simply selecting the highest activity in the saliency map.

The visual attention model of Itti et al. [31] is inspired by the local centre-surround competition mechanisms that account for the non-classical receptive field properties of neurons in the primary visual cortex. An iterative filtering and half-wave rectification scheme similar to a winner-take-all mechanism limits the total number of active sites within each feature map, and these are summed up to produce the saliency map on which a WTA makes the final selection.

One of the problems encountered by computational models that use saliency maps is that they translate physical properties of the stimulus such as luminance, colour, size, onset, etc. into saliency values. Since the various stimulus dimensions have very different characteristics, combining them is a non-trivial problem. The four main approaches presented in the literature are reviewed in [32]: simple normalized summation, linear combination with learned weights, global non-linear normalization followed by summation, and local non-linear competition between salient locations. The study concludes that the best overall results can be obtained by the last two methods, which happen to be simplified versions of what is thought to be biological within-feature spatial competition for saliency.

Draper’s critical evaluation of the Itti et al. model starts from the assumption (questionable from a biological perspective) that attention is simply a front-end for object recognition [33]. If this is the case, the authors conclude that the attentional system must have similar behaviour when faced with transformations of the input, i.e. it must be insensitive to affine transformations of the stimuli in the image. Their analysis finds the model to be lacking, and propose a fairly simple and intuitive solution: instead of a single master saliency map, they use one saliency map for every scale, with global competition that ensures that the system as a whole is insensitive to scaling, rotation and translation of stimuli in the input image (similar to [34]).

The models presented above all use low-level features in the process of building the saliency map. A different approach is taken by Lee et al. [35] with very impressive results, but at the expense of generality. Their Interactive Spiking Neural

Network (ISNN) is geared specifically at finding human faces, and in order to accomplish this they use domain-specific intermediate-level features such as ellipses, aspect-ratio and symmetry, in combination with skin-colour detection. These intermediate level features are combined into a saliency map using binary set operations, each possible combination of features being given a weight through an original learning algorithm. Another interesting aspect of this model is the fact that it does not employ a hierarchical processing scheme, this being a saliency map model in its purest form.

The search for a neural correlate to the concept of a saliency map has led to much interesting experimental work each providing evidence for one or another particular location (superior colliculus [36–38]; LGN [39,40]; V1 [41]; V1 and V2 [42]; pulvinar [43–45]; FEF [46]; parietal [47]). In each of these, the correlate is found by locating maxima of response within a neural population that corresponds to the attended location. This seems to point towards saliency as a distributed computation, and like attention itself, evidence reflecting those computations can be found in many, if not all, neural populations, a point of view taken by a number of models.

In the solution presented in the Selective Tuning model [2], the highest level of the processing hierarchy acts in essence as a very low-resolution saliency map, and it is the attentive mechanism that provides the localization of the attended stimulus through feedback connections that activate pyramids of Θ-winner-take-all[1] competitions. These competitions refine the very coarse initial representation of the attended stimulus, while at the same time pruning connections that interfere with representation of the selected area. A logical consequence of this approach is that each level acts in effect as a saliency map for the features it represents, and thus selection does not need to happen only at the highest level of the pyramid, attention can be directed to any feature that is represented at any level of the feature pyramid.

The Neurodynamical Model [48] implicitly codes saliency as a distribution of modulation across the feature maps. Feature maps relevant for the task are enhanced and/or distracters are inhibited, the dynamics of the network producing winners without the need for explicit representation of salience. This model is split along the temporal/dorsal divide, with the temporal pathway implementing invariant object recognition and the dorsal pathway representing space. The dorsal pathway represents a biasing map with a dual role. In spatial selection mode, a location in this map is selected, and this selection is used to bias the competition in V1 in favour of features located in the corresponding position, features that will be processed first by the object recognition subsystem. In object recognition mode, once a set of features emerges as winner of the distributed competition, the corresponding area in the dorsal pathway is selected, again biasing processing in favour of that spatial location. This biasing map is further used in modelling the symptoms of neglect by introducing a bias gradient in the representation of space, with the various manifestation of the disease associated with different gradient profiles.

4.3. Top-down influences

Many researchers have included top-down influence in their models, here we will review a few characteristic approaches. Two general trends have emerged: in saliency map based models, the top-down influences act directly on the saliency map, making it more likely that targets within certain spatial areas will be selected, while in models based on distributed competition, the competition itself is biased in favour of task relevant stimuli. It is important to note that the two approaches are not incompatible.

Designed specifically as a word recognition system, MORSEL [49] integrates top-down influences in a task-specific fashion. The top-down component of the attentional mechanism moves the focus of attention based on specific information such as static target expectations or dynamic scanning patterns for reading.

As seen above, in the Selective Tuning model [2], due the fact that the model makes a clear distinction between saliency and localization, top-down influences can be integrated in a simple and natural fashion. The authors demonstrate the power of this approach by implementing external biases for or against spatial regions and/or feature maps. Unfortunately, the very simplistic processing pyramid and model neurons used in this first implementation do not allow a comparison between the performance of the model and that of primates, which is the ultimate test for any model of the kind reviewed here. Only scan paths, i.e. qualitative evaluations of the system's performance, are presented, and not reaction times in search tasks.

A model aimed specifically at the aspect of learning top-down influences [50] uses a back-propagation network that receives task-dependent input and controls the flow of information from the low-level feature maps to the saliency (or 'priority') map. The network is trained to enhance the relevant and suppress the irrelevant information for the current task. As shown by [32], this approach can have good results in dealing with specific problems at the expense of generality.

In the Neurodynamical Model of visual attention [48], a sequence of parallel 'where' and 'what' pathways that operate at different spatial resolutions and speeds is presented. The whole system acts as a hierarchical predictor, where the low-resolution analysis determines areas of interest that are investigated in a serial fashion at increasingly higher resolutions, under the guidance of the attentional system.

A number of modellers have decided to either ignore all the other issues or rely purely on mathematical models or on existing models of attentional selection and focus exclusively on the issue of top-down control.

One clear example of this approach is Oliva et al. [51], which is a spin-off from previous work by the same researchers on rapid scene categorization (e.g. [52]), which in turn follows the pioneering work of Biederman [53]. The key point of this previous research has been to demonstrate that low-resolution

[1] Our terminology. In Θ-winner-take-all, units only compete if they differ by more than a threshold Θ.

information is sufficient to categorize a scene, and that this is done very quickly in the brain. The model uses this information to bias or cue a saliency map, and the resulting system is shown to demonstrate very human-like sequences of fixations in natural scenes. While it is questionable that this type of cueing constitutes ‘top-down’ influences, the approach has significant merit, as it and [48] are the first models to explicitly take into account the fact that information is processed at different speeds, and to suggest potential ways in which this phenomenon could be used to influence behaviour.

A similar approach is taken by Lee et al. [35], but, significantly, in this model the quick and dirty processing is combined with true top-down influences, presented to the system in terms of instructions of the form ‘near red’ or ‘above blue’. This is an approach somewhat reminiscent of the concept of ‘indirect search’, introduced and analysed formally by Wixson [54].

4.4. Attentional selection and filtering

Experimental results show that primates can attend to locations, objects or features in the visual field, and within each category, the actual shape and extent of the attentional focus is subject to experimental manipulation.

Most computational models of visual attention include some form of spatial attention, but the shape and size of the attended location and any limitations in this respect are not explicitly addressed in a rigorous fashion. The Itti et al. model [31] seems to assume that attention is a circular ‘spotlight’ of fixed size that just indicates the general area of interest. Approaches that rely on the dynamics of the neural networks for selection, such as [2,55,56], do not make any assumptions about the shape, size or even number of attended areas. While the models do not make any such assumptions, some contiguity criteria are introduced in the implementation of the systems, mainly in the form of Θ-winner-take-all competitions with proximity biases.

In general, it is difficult for modellers to approach the issue of the fate of objects selected (or not selected) by attention since little is known about the high level cortical mechanisms that use the information generated by the object recognition systems of the brain, touching on notions that have so far eluded our understanding, such as consciousness and awareness.

The fate of items not selected by attention is generally not discussed explicitly by modellers, with the notable exception of ST [57]. The fundamental theoretical assumption behind ST is that the role of attention is to eliminate the interference between the stimuli that fall within the receptive fields of neurons, especially at high levels of the visual hierarchy where these can cover large portions of the visual field. The competition between stimuli occurs in all the levels of the hierarchy, guided by top-down influences that in effect bias the competition in favour of the stimuli that are part of (or consistent with) the winning object at the top of the hierarchy. This means that stimuli close to the focus of attention will be inhibited strongly, while stimuli outside this area of inhibition will pass unaltered. This prediction of the model has received significant experimental support [58–71].

Attentional control has not been the focus of intense modelling research, in many cases taking the form of unspecified external mechanisms, e.g. in the Neocognitron system [72] external switch signals disengage attention and allow it to focus on the next target. In some models, the time course of the attentional process is determined by the competitive dynamics of the system [2,55,56]. For example, in the Selective Tuning model [2] the WTA processes generate ‘gating control signals’ responsible for propagating the competition in the appropriate sequence and for determining the duration of one attentional fixation.

Rybak et al. [73] present a model that is centred around the notion of attentional control. In their system, objects are internally represented as sequences of fixations and associated percepts, and a ‘behavioural programme’ selects fixation targets based on hypotheses about the interpretation of the current field of view.

4.5. Inhibition of return, covert and overt attention

While many models of visual attention include demonstrations of inhibition of return (IOR) and overt attention, they are in general based on engineering solutions that have little if anything to do with the way the brain accomplishes this complex task. The fact that biologically plausible implementations are not readily available is a reflection of the fact that little is known about the underlying neural mechanisms and representations involved, rather than a weakness of the models. The experimental evidence for the neural mechanisms involved in IOR is reviewed by Klein [24].

The main reason as to why IOR is almost religiously included in all models is that any system that uses a winner-take-all scheme for selection must have a method for removing the winner from consideration or else the algorithms are stuck and will not find the second strongest and so on. But there is no evidence for the same rationale in the brain. The second reason for IOR, in brain or computer systems, is as a mechanism for adjusting priorities of selection.

For IOR in covert attention, two types of solutions have been presented in the literature. Many models either inhibit the selected locations (e.g. in the saliency map [31]) or inhibit the whole neural pathway corresponding to the selected stimulus [2]. In some cases, the inhibition decays in time, allowing for the locations to be reselected after a while. The second approach, exemplified by the Neocognitron system [72], is to simulate neural fatigue that prevents previously selected neurons from becoming active.

Of course, these simple approaches are not sufficient in active vision systems or in systems that attend to moving targets, where the attended feature’s location changes in time. These cases require higher-level representations of objects and events, representations of the spatio-temporal world in which the visual stimuli appear, mappings from retinal to world-based coordinate systems, and some form of short-term spatial memory. Without these, oscillatory fixations will naturally occur. One example of an attempt to solve this is by Backer and Mertsching [74] who propose a short-term, limited capacity memory of object files. Inhibition of return is implicit, and follows from the fact that object files are assigned priorities based on the time when they were last selected. A review of the issues of coordinate transforms and/or dynamic remapping can be found in [75].

4.6. Attention and recognition

Many modelling efforts separate attention and recognition, and even today, some researchers persist in this approach, e.g. [33] declares that the purpose of a model of visual attention is to be 'the front end to an appearance based object recognition system'.

A clear example of this approach is the Shifter Circuits model [76,77], a model that basically consists of a set of control neurons that dynamically route information from a window on the input to higher areas. Once an area of the input image has been selected, it is presented to an associative recognition network after being transformed to a canonical pose. A similar approach is taken by the SCAN model [78] and by the Selective Attention for Identification Model (SAIM) [79].

MORSEL [49] integrates attentional selection into an object recognition network (in particular, stylised printed words), thus achieving the goal of multi-object processing. The retinal input is processed by a recognition network (called BLIRNET) that maps the raw stimuli to representations of words and letters. With several words in the input image, a simple word recognition system will sometimes incorrectly combine letters to form words that are not present (similar to the 'letter migration' phenomenon observed in perceptual studies [80]). The addition of a separate attentional module is able to overcome this problem. Two distinct attentional selection mechanisms are presented: a late selection component (a 'pull-out net') and an early selection component (the 'attentional mechanism'). The late-selection mechanism acts on the outputs of the recognition network. The attentional mechanism builds a spotlight by combining bottom-up information, biasing selection towards locations that contain input, and top-down task specific information such as static target expectations or dynamic scanning patterns for reading. Note that the selection is not binary, and even non-attended locations get a certain degree of processing.

Another approach that separates attention from object recognition is presented by Walther et al. [81]. In this case, the saliency-based attentional system of Itti et al. [31] operates in parallel to the hierarchical recognition system of Riesenhuber and Poggio [82], and the result of the WTA competition on the saliency map is used as a modulation mask in the layers that represent features of intermediate complexity in the recognition hierarchy. The system seems to work well for simple, paper-clip type objects, but because saliency based on simple features is used in segmentation, in natural images where objects are not uniform in their most salient feature, the system has limitations.

The diametrically opposite approach is the total integration of attention and object recognition, a solution pioneered by Fukushima's Neocognitron system [72]. While the Neocognitron pattern recognition architecture has undergone significant evolution, the form under which attention has been integrated is based on a hierarchical pyramid of simple and complex cells that are trained through unsupervised learning. The last layer of the system, the recognition layer, projects feedback towards the lower layer of the system. Since the feedback signals are gated by the feed-forward pathway, they follow the same route as the feed-forward signals. If a feature is missing, the feedback is blocked, which causes a lowering of the detection threshold in the feed-forward pass, so as to detect even attenuated traces of the input, and the feedback signal continues. This process is repeated until a perfect output is found, the system working in effect as an associative memory. To ensure that only one output is active at any given time, the output layer has lateral inhibitory connections.

Two different schemes of integrated attention and object recognition are investigated in the context of the Neurodynamical model. The first one, presented in [48] has been discussed in Section 4.3. In [56], attention is implemented through local competition biased by top-down connections, while object recognition is implemented in the feed-forward connections that are trained through Hebbian learning. Parallel and somewhat similar structures for invariant object recognition and spatial location are presented, and this allows for a similar treatment of both spatial and object-based top-down influences, manifested by the biasing of the appropriate top-level representations, biases that travel through the network to simulate visual search and object recognition.

Rybak et al. [73] present an object recognition system based on trans-saccadic integration. At each fixation, a local weak classifier is extracted in the form of groups of edges represented in a local coordinate system, and a hypothesis is generated about the identity of the viewed object. This hypothesis is then used by a 'behaviour programme' to generate saccades, and the new views used to validate the hypothesis.

Tsotsos et al. [55] presents an extension of the Selective Tuning model [2] that is able to recognize and localize basic motion patterns in natural image sequences. In this system, high-level motion patterns such as translation, rotation, spiral motion, shear are built up from low level optic flow information and intermediate level motion gradients. Attention selects a winning high-level pattern, and the Selective Tuning feedback process refines its representation and localizes the pattern in the input image sequence.

5. The selective tuning model of visual attention

This section will focus on one particular model, the Selective Tuning (ST) model, providing more detail as well as highlighting how it deals with the issues discussed in the previous chapter. To begin with, and in contrast to any of the other models, ST features a theoretical foundation of provable properties based on the theory of computational complexity

[4–6,13]. The 'first principles' arise because vision is formulated as a search problem (given a specific input, what is the subset of neurons that best represent the content of the image?) and complexity theory is concerned with the cost of achieving solutions to such problems. This foundation suggests a specific biologically plausible architecture as well as its processing stages as will be briefly described in this article (a more detailed account can be found in [2,6]).

5.1. The model

The visual processing architecture is pyramidal in structure with units within this network receiving both feed-forward and feedback connections. When a stimulus is presented to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid with receptive fields (RFs) mapping to the stimulus location; the result is a diverging cone of activity within the processing pyramid. It is assumed that response strength of units in the network is a measure of goodness-of-match of the stimulus within the receptive field to the model that determines the selectivity of that unit.

Selection relies on a hierarchy of winner-take-all processes. WTA is a parallel algorithm for finding the maximum value in a set. First, a WTA process operates across the entire visual field at the top layer where it computes the global winner, i.e. the units with largest response. The fact that the first competition is a global one is critical to the method because otherwise no proof could be provided of its convergence properties. The WTA can accept guidance to favour areas or stimulus qualities if that guidance is available but operates independently otherwise. The search process then proceeds to the lower levels by activating a hierarchy of WTA processes. The global winner activates a WTA that operates only over its direct inputs to select the strongest responding region within its receptive field. Next, all of the connections in the visual pyramid that do not contribute to the winner are pruned (inhibited). The top layer is not inhibited by this mechanism. However, as a result, the input to the higher-level unit changes and thus its output changes. This refinement of unit responses is an important consequence because one of the important goals of attention is to reduce or eliminate signal interference [6]. By the end of this refinement process, the output of the attended units at the top layer will be the same as if the attended stimulus appeared on a blank field. This strategy of finding the winners within successively smaller receptive fields, layer by layer, in the pyramid and then pruning away irrelevant connections through inhibition is applied recursively through the pyramid. The end result is that from a globally strongest response, the cause of that largest response is localized in the sensory field at the earliest levels. The paths remaining may be considered the pass zone of the attended stimulus while the pruned paths form the inhibitory zone of an attentional beam. The WTA does not violate biological connectivity or relative timing constraints. Fig. 1 gives a pictorial representation of this attentional beam.

An executive controller is responsible for implementing the following sequence of operations for visual search tasks:

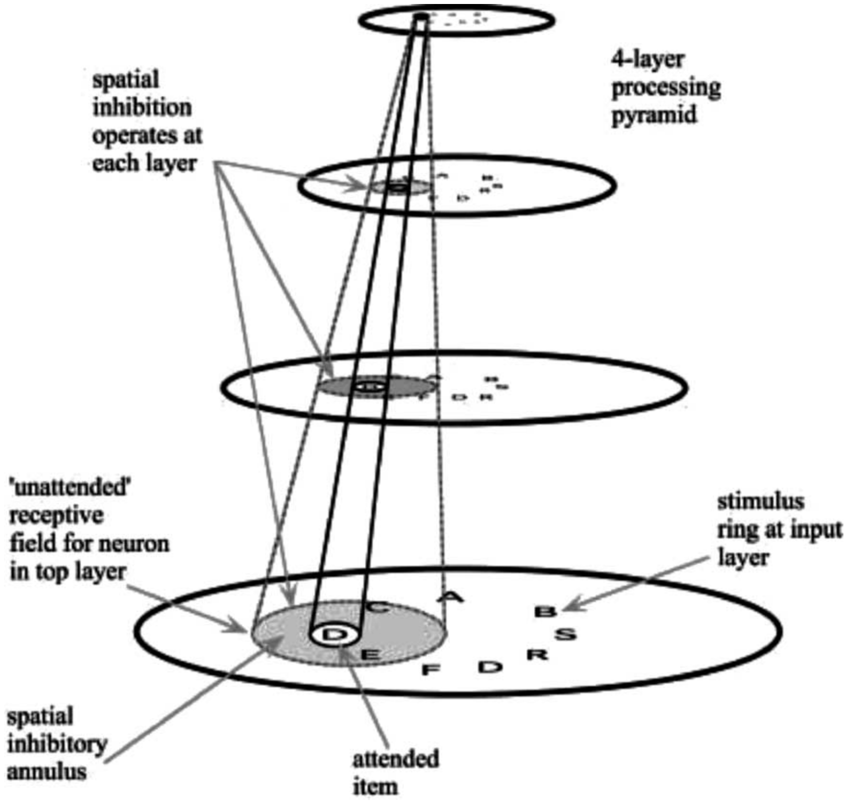
4-layer processing pyramid
spatial inhibition operates at each layer
'unattended' receptive field for neuron in top layer
stimulus ring at input layer
A
B
C
S
D
E
R
F
D
spatial inhibitory annulus
attended item

Fig. 1. Attentional beam. This shows the rationale for suppression around attended items that is a feature of ST.

1. Acquire target as appropriate for the task, store in working memory
2. Apply top-down biases, inhibiting units that compute task irrelevant quantities
3. 'See' the stimulus, activating feature pyramids in a feed-forward manner
4. Activate top-down WTA process at top layers of feature pyramids
5. Implement a layer-by-layer top-down search through the hierarchical WTA based on the winners in the top layer
6. After completion, permit time for refined stimulus computation to complete a second feed-forward pass. Note that this feed-forward refinement does not begin with the completion of the lowermost WTA process; rather, it occurs simultaneously with completing WTA processes (step 5) as they proceed downwards in the hierarchy. On completion of the lowermost WTA, some additional time is required for the completion of the feed-forward refinement.
7. Extract output of top layers and place in working memory for task verification
8. Inhibit pass zone connections to permit next most salient item to be processed
9. Cycle through steps 4–8 as many times as required to satisfy the task.

This multi-pass process may seem to not reflect the reality of biological processes that seem very fast. However, it is not claimed that all of these steps are needed for all tasks. Several different levels of tasks may be distinguished [83], defined as:

Detection is a particular item present in the stimulus, yes or no?
Localization detection plus accurate location;
Recognition localization plus accurate description of stimulus;
Understanding recognition plus role of stimulus in the context of the scene.

The executive controller is responsible for the choice of task based on instruction. If detection is the task, then the winner after step 4, if it matches the target, will suffice and the remaining steps are not needed. Thus, simple detection in this framework requires only a single feed-forward pass. If a localization task is required, then all steps up to 7 are required because, as argued below, the top-down WTA is needed to isolate the stimulus and remove the signal interference from nearby stimuli. This clearly takes more time to accomplish. If recognition is the task, then all steps, and perhaps several iterations of the procedure, are needed in order to provide a complete description. The understanding task has similar requirements, although this is not quite within the scope of the model at this point.

5.2. Top-down selection

Selective Tuning features a top-down selection mechanism based on a coarse-to-fine WTA hierarchy. Why is a purely feed-forward strategy not sufficient? There seems to be no disagreement on the need for top-down mechanisms if task/domain knowledge is considered, although few non-trivial schemes seem to exist. Biological evidence, as well as complexity arguments, suggests that the visual architecture consists of a multi-layer hierarchy with pyramidal abstraction. One task of selective attention is to find the value, location and extent of the most 'salient' image subset within this architecture. A purely feed-forward scheme operating on such a pyramid with:

1. fixed size receptive fields with no overlap, is able to find the largest single input with local WTA computations for each receptive field but location is lost and extent cannot be considered
2. fixed size overlapping receptive fields, suffers from the spreading winners problem, and although the largest input value can be found, the signal is blurred across the output layer, location is lost and extent is ambiguous
3. all possible RF sizes in each layer, becomes intractable due to combinatorics.

While Case 1 might be useful for certain computer vision detection tasks, it cannot be considered as a reasonable proposal for biological vision because it fails to localize targets. Case 3 is not plausible as it is intractable. Case 2 reflects a biologically realistic architecture, yet fails at the task of localizing a target. Given this reality, a purely feed-forward scheme is insufficient to describe biological vision. Only a top-down strategy can successfully determine the location and extent of a selected stimulus in such a biologically realistic architecture.

5.3. WTA and saliency

The Winner-Take-All scheme within ST is defined as an iterative process that can be realized in a biologically plausible manner insofar as time to convergence and connectivity requirements are concerned. The basis for its distinguishing characteristic comes from the fact that it implicitly creates a partitioning of the set of unit responses into bins of width determined by a task-specific parameter, Θ. The partitioning arises because inhibition between units is not based on the value of a single unit but rather on the absolute value of the difference between pairs of unit values. Further, this WTA process is not restricted to converging to single points as all other formulations. The winning bin of the partition is claimed to include the strongest responding contiguous units in the image (this is formally proved in [2]).

A second phase of competition depends linearly on the topographical distance between units, i.e. the features they represent. The larger the distance between units is, the greater the inhibition. This strategy will find the largest, most spatially contiguous subset within the winning bin. A spatially large and contiguous region will inhibit a contiguous region of similar response strengths but of smaller spatial extent because more units from the large region apply inhibition to the smaller region than inhibit the larger region from the smaller one. At the top layer, this is a global competition; at lower layers, it only takes place within receptive fields. In this way, the process does not require implausible connectivity lengths. For efficiency reasons, this is currently only implemented for the units in the winning bin. With respect to the weighted sums computed, in practice the weights depend strongly on the types of computations the units represent. There may also be a task-specific component included in the weights. Finally, a rectifier is needed for the whole operation to ensure that no unit values go below zero. The iterative update continues until there is only one bin of positive response values remaining and all other bins contain units whose values have fallen below Θ. Note that even the winning bin of positive values must be of a value greater than some threshold in order to eliminate false detections due to noise.

It is not assumed that there is a single top-level representation for the WTA process. As a result, there is no single saliency map in this model as there is in most other models. Indeed, there is no single WTA process necessarily, but several simultaneous WTA threads. Saliency is a dynamic, local, distributed and task-specific determination and one that may differ even between processing layers as required. Although it is known that feature combinations of high complexity do exist in the higher levels of cortex, the above does not assume that all possible combinations must exist. Features are encoded separately in a pre-defined set of maps and the relationships of competition or cooperation among them provide the potential for combinations. This flexibility allows for a solution (at least in part) to the binding issues that arise for this domain.

The model has been implemented and tested in several labs applying it to goal computer vision and robotics tasks. The current model structure is shown in Fig. 2. The executive controller and working memory, the motion

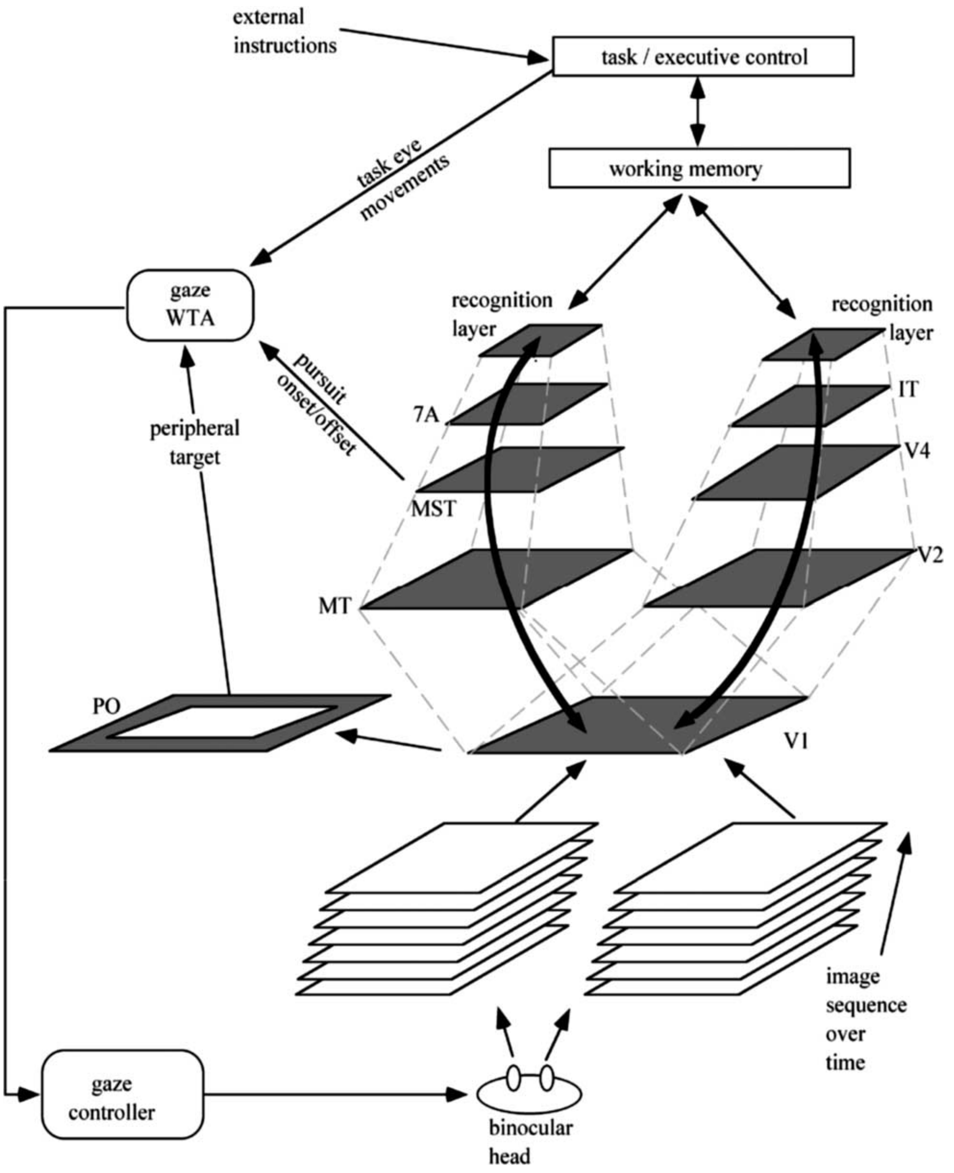

Fig. 2. ST full hierarchy. The full visual processing hierarchy on which ST operates is depicted. This paper focuses on the motion pathway—areas V1, MT, MST, and 7a. Several other components have been demonstrated previously while others are current research topics.

pathway (V1, MT, MST, 7a), the peripheral target area PO, the gaze WTA and gaze controller have all been implemented and examples of performance can be found in [2,3,34,55,84].

5.4. Feature binding

A major contribution of ST is its method of grouping features, demonstrated within a complex motion hierarchy [3,55], known as the binding problem in computational neuroscience [21]. It is not claimed that the particular strategy has sufficient generality to solve all possible issues within the binding problem; however, it seems to solve the limited cases that occur in image sequences of simple motion patterns. As such, it is the first instance of such a solution and further work will investigate its generality.

Following Roskies [21], 'the canonical example of binding is the one suggested by Rosenblatt […] in which one sort of visual feature, such as an object's shape, must be correctly associated with another feature, such as its location, to provide a unified representation of that object.' Such explicit association ('binding') is particularly important when more than one visual object is present, in order to avoid incorrect combinations of features belonging to different objects, otherwise known as 'illusory conjunctions' [85]. Several other examples of the varieties of binding problems in the literature appear in *Neuron* 24(1) (1999). At least some authors [86,87] suggest that specialized neurons that code feature combinations (introduced as cardinal cells by Barlow [88]) may assist in binding. The ST solution does indeed include such cells; however, they do not suffice on their own as will be described because they alone cannot solve the localization problem.

Using the classical view of the binding problem, it is straightforward to show that for a purely data-directed strategy, the problem of finding the subsets of each feature map that correspond to the parts of an object has exponential complexity (it is an instance of the NP-Complete visual matching problem [5]). In simple detection problems, the complexity is manageable by simple strategies because there are not too many choices and the task is simply detection of a target. However, in the general case, a top-down attentional selection mechanism is needed to reduce the complexity of the search problem. It is for this reason that attention constitutes the link between sensing and recognition.

The use of localized saliency and WTA decision processes is precisely what the binding problem requires: neurons in different representations that respond to different features and in different locations are selected together, the selection being in location and in feature space, and are thus bound together via the 'pass' zone(s) of the attention mechanism. Even if there is no single neuron at the top of the pyramid that represents the concept, the WTA allows for multiple threads bound through location, as shown earlier.

Part of the difficulty facing research on binding is the confusion over definitions and the wide variety of tasks included in binding discussions. For example, in Feature Integration Theory (FIT) [89], location is a feature because FIT assumes location is faithfully represented in a master map of locations. But this cannot be true; location precision changes layer to layer in any pyramid representation. In the cortex, it is not accurate in a Euclidean sense almost anywhere, although the topography is qualitatively preserved [90]. The wiring pattern matters in order to get the right image bits to the right neurons. Thus, binding needs to occur layer to layer and is not simply a problem for high-level consideration. Features from different representations with different location coding properties converge onto single cells and this seems to necessitate an active search process.

The key point in ST's approach to binding is that location is not a feature, rather, it is the anchor that permits features to be bound together. Here, location is defined broadly and differently in each area of the visual hierarchy and in practice is considered to be local within a visual area—e.g. an array of hypercolumns, each with its own local coordinates. If a grouping of features is not coincident by location, it can only be considered as a unitary group if they converge onto units representing the group, and features that compose a group may be in different locations and represented in different visual areas as long as this convergence criterion is met. If a group is attended, then the WTA described earlier will find and attend to each of its parts regardless of their location or feature map representation. It is the WTA mechanism, guided by task information that decides which groups of features are compatible with each other and with the selected representation, and eliminate the incompatible ones [3].

This strategy is sufficient to handle complex recognition task such as multiple patterns and overlapping objects (see [3] for examples). As such, it is a solution to the aspect of binding that attends to groups and finds parts of groups.

6. Discussion

The preceding sections have presented arguments that explicit strategies for visual attentional selection are important for cognitive vision systems, and have moreover shown that a number of proposals currently exist for exactly how parts of this goal may be accomplished. A key characteristic of virtually all the models surveyed above is that they receive significant inspiration from the neurobiology and psychophysics of human and primate vision. This, although not necessarily a key component of mainstream computer vision, seems very appropriate for cognitive vision systems given the definition of the topic given in Section 1.

In this review, we have discussed the fact that attention is not a unitary concept, but a series of related phenomena, and we have identified and classified the main questions that a unified theory of attention needs to answer. Within this framework, we analysed current biologically plausible computational models of visual attention, identifying historic trends and major open questions. Our own research in the context of the Selective Tuning model was discussed, highlighting recent results.

The true test of any theory or model lies in its ability to generate verifiable predictions, and ST is one model where a number of predictions, most first presented in 1990, have received significant support:

- An early prediction [6] was that attention seems necessary at any level of processing where a many-to-one mapping of neurons was found. Further, attention occurs in all the areas in concert. The prediction was made at a time when good evidence for attentional modulation was known for area V4 only [91]. Since then, attentional modulation has been found in many other areas both earlier and later in the visual processing stream, and that it occurs in these areas simultaneously [92]. Evidence cited by Britten [93] who reached the conclusion that 'attention is everywhere', save for the Moran and Desimone work, was all post-1990. Vanduffel et al. [71] and O'Connor et al. [94] have shown that attentional modulation appears at early as the LGN.
- Another early prediction of ST is that attentional modulation in higher areas precedes that in V1, prediction supported (at least in the ventral pathway) by Mehta et al. [95]. In general, the model claims that the latency of attentional modulations *decreases* from lower to higher visual areas.
- The notions of competition and of attentional inhibition were also early components of the model [6] and this too has gained evidence over the years [14,92,96].
- The model has always included an inhibitory surround component [6]. This implies that perception may be negatively affected in the vicinity of the attended stimulus. This too has gained support [58–71].
- The model also explains how so-called pre-attentive vision is only a special case of attentive processes [6]; no separate pre-attentive process operates independently of attention, a view Joseph et al. [27] seem to be suggesting too.

How can an attentional selection system be integrated into a cognitive computer vision system? It is certainly true that most if not all such systems have some early vision processing stages. These models, particularly ST, provide a skeleton within which one can include layers of early vision filters and organize them into meaningful hierarchies. It is also true that somewhere in the processing stages, the need to segment an image into regions or events is important. ST's selection strategy may assist with this. If a target object is specified in advance, ST can be shown this in advance, that particular image can be processed and then stored in working memory, and used to guide visual search.

It is clear that ST can provide selection of visual field, selection of detailed sub-regions for analysis, selection of spatial and feature dimensions of interest, and selection of parameters for low-level operations. It cannot, nor can any other of the models surveyed above, select relevant objects or events from a knowledge base with respect to a particular task, select tasks relevant for a domain, select the world model appropriate for solving the current task, and so on. In other words, the machinery described seems appropriate for early and intermediate levels of visual processing but has not yet advanced to stage to be as useful for higher levels of visual processing or for the task levels of processing. These must remain topics for future research.

Acknowledgements

This research was funded by grants from NSERC and NCE IRIS to JKT. ALR is partially funded by NSERC. JKT holds the Canada Research Chair in Computational Vision. The authors thank the two anonymous reviewers for their comments and suggestions.

References

[1] D. Vernon, A vision on cognitive vision, in: Dagstuhl Seminars, 2003
[2] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, F. Nuflo, Modeling visual-attention via selective tuning, Artificial Intelligence 78 (1–2) (1995) 507–545.
[3] J.K. Tsotsos, Y. Liu, J.-C. Martinez-Trujillo, M. Pomplun, E. Simine, K. Zhou, Attending to motion. Computer Vision and Image Understanding, 100(1–2) (2005) 3–40.
[4] J.K. Tsotsos, A 'complexity level' analysis of vision, in: International Conference on Computer Vision: Human and Machine Vision Workshop, London, England, 1987
[5] J.K. Tsotsos, The complexity of perceptual search tasks, in: International Joint Conference on Artificial Intelligence, Detroit, 1989
[6] J.K. Tsotsos, Analyzing vision at the complexity level, Behavioral and Brain Sciences 13 (3) (1990) 423–444.
[7] H.V. Helmholtz, J.P.C. Southall, Helmholtz's Treatise on Physiological Optics, 3, The Optical Society of America, Rochester, NY, 1924.
[8] J.E. Hoffman, in: H. Pashler (Ed.), Visual Attention and Eye Movements, University College London Press, London, 1998, pp. 119–154. Attention.
[9] J. Aloimonos, Active perception, Computer Vision, Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1993. p. 292, viii.
[10] G. Sandini, F. Gandolfo, E. Grosso, M. Tistarelli, Vision during action, in: J. Aloimonos (Ed.), Active Perception, Lawrence Erlbaum Associates, London (Hillsdale, NJ), 1993.
[11] S. Baluja, D. Pomerleau, Dynamic relevance: vision-based focus of attention using artificial neural networks, Artificial Intelligence 97 (1–2) (1997) 381–395.
[12] E.D. Dickmanns, A general dynamic vision architecture for UGV and UAV, Journal of Applied Intelligence 2 (1992) 251–270.
[13] J.K. Tsotsos, On the relative complexity of active vs passive visual-search, International Journal of Computer Vision 7 (2) (1992) 127–141.
[14] R. Desimone, J. Duncan, Neural mechanisms of selective visual-attention, Annual Review of Neuroscience 18 (1995) 193–222.
[15] V. Maljkovic, K. Nakayama, Priming of pop-out: I. Role of features, Memory and Cognition 22 (6) (1994) 657–672.
[16] M.M. Chun, Y. Jiang, Contextual cueing: implicit learning and memory of visual context guides spatial attention, Cognitive Psychology 36 (1) (1998) 28–71.
[17] J.M. Wolfe, Moving towards solutions to some enduring controversies in visual search, Trends in Cognitive Sciences 7 (2) (2003) 70–76.
[18] J.M. Wolfe, N. Klempen, K. Dahlen, Postattentive vision, Journal of Experimental Psychology Human Perception and Performance 26 (2) (2000) 693–716.
[19] S. Yantis, J. Schwarzbach, J.T. Serences, R.L. Carlson, M.A. Steinmetz, J.J. Pekar, S.M. Courtney, Transient neural activity in human parietal cortex during spatial attention shifts, Nature Neuroscience 5 (10) (2002) 995–1002.
[20] S.P. Tipper, J. Driver, Negative priming between pictures and words in a selective attention task: evidence for semantic processing of ignored stimuli, Memory and Cognition 16 (1) (1988) 64–70.
[21] A.L. Roskies, The binding problem, Neuron 24 (1) (1999) 111–125.
[22] R.A. Rensink, When good observers go bad: change blindness, inattentional blindness, and visual experience, Psyche (2000) 6.

[23] J.K. Tsotsos, Triangles, Pyramids, Connections and Attentive Inhibition, PSYCHE: An Interdisciplinary Journal of Research on Consciousness, 1999(http://psyche.cs.monash.edu.au/v5/psyche-5-20-tsotsos.html)

[24] R.M. Klein, Inhibition of return, Trends in Cognitive Sciences 4 (4) (2000) 138–147.

[25] T.S. Horowitz, J.M. Wolfe, Visual search has no memory, Nature 394 (6693) (1998) 575–577.

[26] A. Ignashchenkova, P.W. Dicke, T. Haarmeier, P. Thier, Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention, Nature Neuroscience 7 (1) (2004) 56–64.

[27] J.S. Joseph, M.M. Chun, K. Nakayama, Attentional requirements in a 'preattentive' feature search task, Nature 387 (6635) (1997) 805–807.

[28] M.B. Ben-Av, D. Sagi, J. Braun, Visual attention and perceptual grouping, Perception & Psychophysics 52 (3) (1992) 277–294.

[29] J. Braun, D. Sagi, Vision outside the focus of attention, Perception & Psychophysics 48 (1) (1990) 45–58.

[30] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology 4 (4) (1985) 219–227.

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

[32] L. Itti, C.A. Koch, Comparison of feature combination strategies for saliency-based visual attention systems, in: SPIE Human Vision and Electronic Imaging-IV, San Jose, CA, 1999

[33] B. Draper, A. Lionelle, Evaluation of selective attention under similarity transforms, in: International Workshop on Attention and Performance in Computer Vision (WACPV 2003), Graz, Austria, 2003

[34] W. Wai, J.K. Tsotsos, Directing attention to onset and offset of image events for eye-head movement control, in: IAPR Conference on Pattern Recognition, Jerusalem, 1994

[35] K.W. Lee, H. Buxton, J.F. Feng, Selective attention for cue-guided search using a spiking neural network, in: International Workshop on Attention and Performance in Computer Vision, Graz, Austria, 2003

[36] G.D. Horwitz, W.T. Newsome, Separate signals for target selection and movement specification in the superior colliculus, Science 284 (1999) 1158–1161.

[37] A.A. Kustov, D.L. Robinson, Shared neural control of attentional shifts and eye movements, Nature 384 (1996) 74–77.

[38] R.M. McPeek, E.L. Keller, Saccade target selection in the superior colliculus during a visual search task, Journal of Neurophysiology 88 (2002) 2019–2034.

[39] C. Koch, A theoretical analysis of the electrical properties of an X-cell in the cat's LGN: does the spine-triad circuit subserve selective visual attention? MIT, Artificial Intelligence Laboratory, 1984

[40] S.M. Sherman, C. Koch, The control of retinogeniculate transmission in the mammalian lateral geniculate nucleus, Experimental Brain Research 63 (1986) 1–20.

[41] Z. Li, A saliency map in primary visual cortex, Trends in Cognitive Sciences 6 (1) (2002) 9–16.

[42] T.S. Lee, C. Yang, R.D. Romero, D. Mumford, Neural activity in early visual cortex reflects behavioral experience and higher order perceptual saliency, Nature Neuroscience 5 (6) (2002) 589–597.

[43] S.E. Petersen, D.L. Robinson, J.D. Morris, Contributions of the pulvinar to visual spatial attention, Neurophysiologia 25 (1987) 97–105.

[44] M.I. Posner, S.E. Petersen, The attention system of the human brain, Annual Review of Neuroscience 13 (1990) 25–42.

[45] D.L. Robinson, S.E. Petersen, The pulvinar and visual salience, Trends in Neuroscience 15 (4) (1992) 127–132.

[46] K.G. Thompson, N.P. Bichot, J.D. Schall, Dissociation of visual discrimination from saccade programming in macaque frontal eye field, Journal of Neurophysiology 77 (2) (1977) 1046–1050.

[47] J.P. Gottlieb, M. Kusunoki, M.E. Goldberg, The representation of visual salience in monkey parietal cortex, Nature 391 (6666) (1998) 481–484.

[48] G. Deco, J. Zihl, A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system, Journal of Computational Neuroscience 10 (3) (2001) 231–253.

[49] M.C. Mozer, The perception of multiple objects: a connectionist approach, Neural Network Modeling and Connectionism, MIT Press, Cambridge, Mass, 1991. p. 217.

[50] P. vandeLaar, T. Heskes, S. Gielen, Task-dependent learning of attention, Neural Networks 10 (6) (1997) 981–992.

[51] A. Oliva, A. Torralba, M.S. Castelhano, J.M. Henderson, Top-down control of visual attention in object detection, in: IEEE International Conference on Image Processing, Barcelona, Spain, 2003

[52] A. Torralba, A. Oliva, Statistics of natural image categories, Network-Computation in Neural Systems 14 (3) (2003) 391–412.

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

[54] L. Wixson, Gaze Selection for Visual Search, University of Rochester Dept. of Computer Science, Rochester, NY, 1994. p. xiv, 155 p.

[55] J.K. Tsotsos, M. Pomplun, Y. Liu, J.C. Martinez-Trujillo, E. Simine, Attending to motion: localizing and labeling simple motion patterns in image sequences, in: Conference on Biologically-Motivated Computer Vision, Tuebingen, Germany, 2002

[56] E.T. Rolls, G. Deco, Computational neuroscience of vision, Oxford University Press, Oxford; New York, 2002. xviii, 569 p.

[57] J.K. Tsotsos, Analyzing vision at the complexity level, Behavioral and Brain Sciences 14 (4) (1991) 768.

[58] F. Cutzu, J.K. Tsotsos, The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item, Vision Research 43 (2) (2003) 205–219.

[59] S.D. Slotnick, J. Schwarzbach, S. Yantis, Attentional inhibition of visual processing in human striate and extrastriate cortex, Neuroimage 19 (4) (2003) 1602–1611.

[60] A. Kristjansson, K. Nakayama, The attentional blink in space and time, Vision Research 42 (17) (2002) 2039–2050.

[61] N.G. Muller, A. Kleinschmidt, The attentional 'spotlight''s' penumbra: center-surround modulation in striate cortex, Neuroreport 15 (6) (2004) 977–980.

[62] J.R. Mounts, Evidence for suppressive mechanisms in attentional selection: feature singletons produce inhibitory surrounds, Perception & Psychophysics 62 (5) (2000) 969–983.

[63] J.R.W. Mounts, R.D. Melara, Attentional selection of objects or features: evidence from a modified search task, Perception & Psychophysics 61 (2) (1999) 322–341.

[64] N.G. Muller, M. Mollenhauer, A. Rosler, A. Kleinschmidt, The attentional field has a Mexican hat distribution, Vision Research 45 (9) (2005) 1129–1137.

[65] J.D. Schall, Neural basis of saccade target selection, Reviews in Neuroscience 6 (1) (1995) 63–85.

[66] J.D. Schall, On the role of frontal eye field in guiding attention and saccades, Vision Research 44 (12) (2004) 1453–1467.

[67] S.D. Slotnick, J.B. Hopfinger, S.A. Klein, E.E. Sutter, Darkness beyond the light: attentional inhibition surrounding the classic spotlight, Neuroreport 13 (6) (2002) 773–778.

[68] D.O. Bahcall, E. Kowler, Attentional interference at small spatial separations, Vision Research 39 (1) (1999) 71–86.

[69] G. Caputo, S. Guerra, Attentional selection by distracter suppression, Vision Research 38 (5) (1998) 669–689.

[70] B.A. Steinman, S.B. Steinman, S. Lehmkuhle, Visual attention mechanisms show a center-surround organization, Vision Research 35 (13) (1995) 1859–1869.

[71] W. Vanduffel, R.B.H. Tootell, G.A. Orban, Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system, Cerebral Cortex 10 (2) (2000) 109–126.

[72] K. Fukushima, T. Imagawa, E. Ashida, Character recognition with selective attention, in: International Joint Conference on Neural Networks, Seattle, 1991.

[73] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, N.A. Shevtsova, A model of attention-guided visual perception and recognition, Vision Research 38 (1998) 2387–2400.

[74] G. Backer, B. Mertsching, Two selection stages provide efficient object-based attentional control for dynamic vision, in: International workshop

on attention and performance in computer vision (WACPV 2003), Graz, Austria, 2003

[75] A. Pouget, T.J. Sejnowski, Dynamical remapping, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, MIT Press, Boston, 1995, p. 335.

[76] C. Anderson, D. Van Essen, Shifter circuits: a computational strategy for dynamic aspects of visual processing, Proc. Natl Acad. Sci. USA 84 (1987) 6297–6301.

[77] B.A. Olshausen, C.H. Anderson, D.C. Van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, Journal of Neuroscience 13 (11) (1993) 4700–4719.

[78] E.O. Postma, H.J. vandenHerik, P.T.W. Hudson, SCAN: a scalable model of attentional selection, Neural Networks 10 (6) (1997) 993–1015.

[79] D. Heinke, G.W. Humphreys, SAIM: a model of visual attention and neglect, in: 7th International Conference on Artificial Neural Networks, Lausanne, Switzerland: Springer, 1997

[80] M.C. Mozer, Letter migration in word perception, Journal of Experimental Psychology. Human Perception and Performance 9 (4) (1983) 531–546.

[81] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch, Attentional selection for object recognition—a gentle way, Biologically Motivated Computer Vision, Proceedings, 2002, 2525, pp. 472–479

[82] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neuroscience 2 (11) (1999) 1019–1025.

[83] J.K. Tsotsos, Behaviorist intelligence and the scaling problem, Artificial Intelligence 75 (2) (1995) 135–160.

[84] S. Culhane, J.K. Tsotsos, An attentional prototype or early vision, The Second European Conference on Computer Vision, Springer Verlag, Santa Margherita Ligure, Italy, 1992.

[85] A. Treisman, H. Schmidt, Illusory conjunctions in the perception of objects, Cognitive Psychology 14 (1) (1982) 107–141.

[86] G.M. Ghose, J. Maunsell, Specialized representations in visual cortex: a role for binding?, Neuron 24 (1) (1999) 79–85 pp. 111–25.

[87] C. von der Malsburg, The what and why of binding: the modeler's perspective, Neuron 24 (1) (1999) 95–104 pp. 111–25.

[88] H.B. Barlow, Single units and sensation: a neuron doctrine for perceptual psychology?, Perception 1 (4) (1972) 371–394.

[89] A.M. Treisman, G. Gelade, Feature-integration theory of attention, Cognitive Psychology 12 (1) (1980) 97–136.

[90] D.J. Felleman, D.C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex, Cerebral Cortex 1 (1) (1991) 1–47.

[91] J. Moran, R. Desimone, Selective attention gates visual processing in the extrastriate cortex, Science 229 (4715) (1985) 782–784.

[92] S. Kastner, P. De Weerd, R. Desimone, L.G. Ungerleider, Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI, Science 282 (5386) (1998) 108–111.

[93] K.H. Britten, Cortical neurophysiology—attention is everywhere, Nature 382 (6591) (1996) 497–498.

[94] D.H. O'Connor, M.M. Fukui, M.A. Pinsk, S. Kastner, Attention modulates responses in the human lateral geniculate nucleus, Nature Neuroscience 5 (11) (2002) 1203–1209.

[95] A.D. Mehta, I. Ulbert, C.E. Schroeder, Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas, Cerebral Cortex 10 (4) (2000) 343–358.

[96] J.H. Reynolds, L. Chelazzi, R. Desimone, Competitive mechanisms subserve attention in macaque areas V2 and V4, Journal of Neuroscience 19 (5) (1999) 1736–1753.