Computational Models of Primate Visual Attention

Albert L. Rothenstein

John K. Tsotsos

Technical Report CS-2004-09

December 10, 2004

Department of Computer Science and Engineering

4700 Keele Street North York, Ontario M3J 1P3 Canada

Computational Models Of Primate Visual Attention

Albert L. Rothenstein[1,2] and John K. Tsotsos[1,2]

[1]Dept. of Computer Science, York University, Toronto, Canada

[2]Centre for Vision Research, York University, Toronto, Canada

Correspondence to albertlr@cs.yorku.ca

Abstract

This paper presents a comprehensive survey of approaches to the computational modeling of visual attention. A key characteristic of virtually all the models surveyed 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. The review is placed in the context of related computer vision and neuroscience research. Major theories of primate visual attention are presented in an original classification.

1. Introduction

1.1. Overview

In this paper we will review the major theories and computational models of primate visual attention.

To put this review in perspective, the paper starts with a definition of visual attention that highlights the fact that attention is a set of strategies that have evolved to minimize the computational load of vision. A number of areas in computer vision that take a similar approach are briefly mentioned, in the context of the general trends in the field.

Because any modeling effort has to be based on a thorough understanding of the natural phenomenon under investigation, the paper continues with a brief presentation of the most important experimental paradigms and findings that have guided the development of computational models of visual attention.

The paper continues by introducing the key components of any complete theory of attention, and based on these, presents the major theories of primate visual attention in an original classification. With this foundation in place, we will present the criteria by which some models have been included while others have not, and analyze the major classes of models of visual attention and how they address the identified key components.

The paper will end with a brief summary and some conclusions about the state of the art in the field, highlighting the major open issues.

1.2. What is visual attention?

In his famous 1890 definition, psychologist William James proposed: “Everybody knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. It implies withdrawal from some things in order to deal effectively with others”. In the century that has passed since, we have reached the point where we can confidently say “no one knows what attention is” [1].

Of course, this rather pessimistic outlook hides a less dramatic reality. While we are no further in defining visual attention, we have made tremendous progress in understanding the nature of attention and its mechanisms. From a computational point of view, the long-standing argument has been that the brain is simply not large enough to process all incoming stimuli, but without a quantitative analysis, this argument is not satisfactory. The theoretical analysis is provided by Tsotsos within the framework of computational complexity [2], while at the same time providing constraints on the processing system. The proof starts by showing that purely data-directed visual search in its most general form is an intractable problem in any realization, and concludes that attentive selection based on task knowledge is a powerful heuristic to limit search and make the overall problem tractable. 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.* Of course, this definition is rather broad, but, as we will see while reviewing the experimental evidence, visual attention manifests itself through a very large spectrum of phenomena, influencing all aspects of vision, so any simplistic, narrow definition is likely to miss important aspects.

1.3. Computer Vision Approaches

The inherent complexity of visual processing has been recognized as a problem from the early days of computer vision, and the history of the field has been, at least partially, a long struggle to find ways to reduce this complexity. Of interest here are methods similar in nature to visual attention, namely methods that attempt to selectively process subsets of an image in the hope that if these areas are carefully selected, sufficient information can be extracted to perform the desired task, be it object recognition, autonomous navigation, image compression, etc. A few broad directions have emerged, which can be classified as (adapted from Tsotsos [3]): artificial manipulations, active vision, perceptual grouping, and region of interest operators.

1.3.1. Artificial manipulations

Artificial manipulations are task-specific engineering solutions, based on assumptions about relevant features, knowledge of task domain, controlled environments, etc. In addition, quite often these types of systems are not purely vision-based, integrating information from other modalities. Due to their somewhat ad-hoc nature (even if they are based on rigorous mathematical concepts and proofs), these methods rarely generalize well beyond their intended domain of application. We will only present one example of such a system, illustrating both the power and limitations of the approach. One of the leaders in biometric solutions, ZN Vision Technologies AG (http://www.zn-ag.com/), produces a face recognition system for controlling access to secure areas, ZN-Face. A person requesting access to the secure area has to stand in front of a camera, and the system will take a snapshot of the person's face and apply an elastic graph-matching algorithm on a large number of facial features. The system requires the person to stand in a predetermined position and pose, under highly controlled illumination. The face is segmented out using an active infrared detector that is also used to make sure what is presented is a live person and not a mask or photograph. While the performance of the system is impressive, yielding very low false positive and false negative rates, it is clear that the approach can not be generalized to other application areas due to the very strong task-specific assumptions built into the solution.

1.3.2. Active vision

Active vision systems extract information from a visual scene by manipulating the parameters of their sensory apparatus – in general one or more cameras mounted with one or more mechanical degrees of freedom in addition to the internal degrees of freedom of the cameras (zoom, focus). The basic ideas and first systems emerged in the late 1980s [4-7]. The key observation is that many traditional vision problems could be solved with simple algorithms by using controlled sensor motion [5]. Many systems are based on the primate occulomotor system, with separate tracking (smooth pursuit) and saccade subsystems. From a control point of view, the most common choice for tracking is the PI (proportional band/integral) controller, embedded within a predictor to deal with time delays – linear predictors (e.g. [8]) or Smith predictors (e.g. [9-12]), or Kalman filters (e.g. [10, 12, 13]), while for saccades, sampled-loop [9, 11] or open-loop controller are used [10, 12, 14].

In terms of application domain, two main directions have emerged: object recognition and structure reconstruction. For object recognition, the approach is to sample an input image through a saccadic search algorithm and encode the resulting sub-images using a spatial (and sometimes spatio-temporal) encoding mechanism. These signatures are then classified or recognized using neural networks or databases of images [15-18]. Krotkov et al. present the key ideas behind active scene structure reconstruction, where the parameters of the system (e.g. focus, vergence) are modified in order to improve the performance of structure-from-stereo algorithms [19]. For stereo matching, in [20, 21] matching is restricted to a short range of disparities close to zero, and then the camera vergence is varied. [22, 23] use a combination of disparity, vergence, focus, and aperture to build scene models over multiple fixations. Foveal multiresolution sensors are used to improve performance in [22, 24, 25]. An interesting aspect of [25] is the ability of their system to efficiently determine correspondences between non-uniform levels of spatial resolution.

1.3.3. Perceptual grouping

Perceptual grouping encompasses processes that attempt to organize image features into structures, thus allowing computer vision systems to move away from pixel intensity data and deal with higher level constructs, allowing symbolic manipulations and other traditional artificial intelligence techniques to be applied. Perceptual grouping, as the link between low-level segmentation and high-level algorithms, has the potential to make significant contributions to figure-ground segregation, object recognition, scene reconstruction, change detection, spatio-temporal grouping, and many other areas. Many approaches fall under the general heading of perceptual grouping, and different groups use the same names to denote different processes, or different names for the same process. This makes comparisons difficult, a problem first identified by Sarkar and Boyer [26], a review paper that also proposed a unified terminology. A decade later Engberts and Smeulders argue that little has changed in this respect [27]. In general, clustering/classifying/feature grouping deal with generic methods that have not been developed specifically for computer vision, such as k-means, graph-theoretic clustering, nearest neighbor, and neural networks, and as such, their application is quite often problem-specific (see [26, 27] for a thorough discussion of these and other issues). Keeping with the biologically inspired theme of this paper, we will define perceptual grouping as being concerned with simulating and understanding the mechanisms that underlie Gestalt grouping principles. The role of these principles in computer vision was demonstrated by [28, 29]. Early work (pre 1993) is reviewed in detail by Sarkar and Boyer [26], where they also propose a classification structure for perceptual organization, and based on this they identify areas of potential future research, especially in the area of texture flows and motion.

Texture flows, defined as two-dimensional structures characterized by local parallelism and slowly varying dominant local orientation, were first approached by Stevens [30], using a histogram based approach on a sparse representation of the flow. The idea of using dense, vector based representations that link local and global structure through differential equations, was introduced by Zucker [31]. [32-34] perform texture flow segmentation using gradient descent to control the smoothing of orientation diffusion. Ben-Shahar and Zucker [35] extend earlier work [31] to handle sparse data sets using a variation on orientation diffusion and relaxation labeling to asses the degree to which a particular data point is consistent with the context in which it is embedded, and whether or not that context is part of a single object or not.

In the motion domain, Little et al. [36] introduce the idea of computing optic flow using a local voting scheme based on similarity of planar patches, an idea that was extended by [37, 38] with the introduction of tensor voting, a method that better preserves discontinuities, while at the same time improving the performance of the algorithms by being non-iterative. In general, many methods suffer from the over-smoothing problem (e.g. the wavelet based approach of [39]), or from a strong dependence on initial conditions (e.g. the recursive partitioning method of [40] or the steerable flow field basis set of [41]).

1.3.4. Region of interest operators

Region of interest operators are primarily used in indexing to “summarize” images for fast querying, in image/object recognition to provide invariant descriptions of important features and in intelligent image compression to guide lossy compression algorithms. Regions of interest are defined by Haralick and Shapiro [42] as points in an image that are distinguishable from their immediate neighbors, and the position as well as the selection of the interesting point should be invariant with respect to the expected geometric and radiometric distortions.

In the next two sections we will review in some detail the area that has the most similarity with the computational modeling of visual attention, namely region of interest operators, with a particular emphasis on biologically inspired approaches.

1.3.4.1. Region of interest operators – Moravec’s legacy

Region of interest operators, originating in the early 1980s in the robotics community, are primarily used in indexing to “summarize” images for fast querying, in image/object recognition to provide invariant descriptions of important features and in intelligent image compression to guide lossy compression algorithms. A region of interest is defined by Haralick and Shapiro [42] as a point in an image that has two main properties: distinctiveness and invariance. This means that a point should be distinguishable from its immediate neighbors, and the position as well as the selection of the interesting point should be invariant with respect to the expected geometric and radiometric distortions.

Moravec’s research is the first recorded use of region of interest operators is [43, 44], where it was used for stereo matching in the control of a mobile robot, the Stanford Cart. Moravec’s operator detects points where intensity changes abruptly in at least one direction. For most of the intervening years, the trend has been towards improving region of interest operators by using better features and better matching heuristics, and towards extending the approach to other types of images. The first logical step was the use of corner detectors, and the Harris detector [45, 46] makes the process more repeatable over image variations and near the edges, while extending the approach to motion tracking and the recovery of 3D structure from motion.

Among the features tried as alternatives to the Harris detector, biologically inspired and statistical approaches stand out. For example, orientation-selective Gabor filters and their first and second derivatives are used in [47], while [48] uses blind estimates of signal-dependent noise variation based on local statistics of pixels classified as region/contour/interest point using autocorrelation.

Other improvements increased the range of changes over which features could be matched. Zhang et al. [49] use correlation windows and outlier removal (majority vote on geometric rigid constraints) to match corners over large changes. A similar approach is taken by Torr [50] for long range motion matching. A major step forward in the field was the introduction of rotationally invariant descriptors in Schmid and Mohr [51], the first work to use region of interest operators in general image recognition. Another important contribution was feature clustering, used to deal with occlusions and clutter. The next logical step is scale-invariance, introduced by Lindeberg [52], Mikolajczyk and Schmid [53] and Lowe [54]. Lowe's novel local descriptor was also less sensitive to local image distortions.

Schmid et al. [55] evaluate the state of the art, and draw two main conclusions for future research: detectors should be included in multi-scale frameworks and in order to generate detectors with high information contents, image statistics need to be studied.

The current trend seems to be towards the use of wavelet transforms [56-59], that seem to be able to extract points with larger information content and provide better repeatability [57]. The caveat is that this work is mainly focused on content-based image retrieval, so the geometric stability of the selected points under transformations in not an issue. It remains to be seen how the approach will fare in the general case.

It can be seen from this brief overview that Moravec's legacy is still strong, but since his goal was to guide indoor mobile robots, this work can not be extended easily to natural images, where corners are rare, most likely to be detected in textured areas, so points of interest will be strongly clustered, not the best thing to do.

1.3.4.2. Attention in computer vision

Of particular interest within the area of region of interest operators are biologically inspired approaches that implement some form of visual attention. These systems do not attempt to model neuroscience and psychology results, but simply take them as starting points in the development of engineering solutions to the problem of the computational complexity of vision. The general approach is to augment some form of region of interest operator with intermediate or high level information, in the form of either structural or semantic information.

Burt [60] provides a theoretical justification for the need for attentional mechanisms in computer vision systems and suggests possible solutions for three subsets of the problem, identified as foveation, tracking and high level interpretation. Under foveation, equivalent to primate saccadic eye movements, older results regarding the value of a non-uniform acuity sensor are reviewed in detail, and a computationally efficient method for representing this kind of data is presented. Tracking is analogous to primate smooth pursuit eye movements, and its purpose is to stabilize an object's moving image on the sensor. Simple feedback control is proposed as a solution, based on multi-resolution image registration. Finally, high level interpretation is used in a hypothesis testing paradigm to direct the high resolution processing to areas of the image that are most likely to contain information relevant for the task. The solution suggested is a symbolic stage composed of "pattern trees" that describe a database of known objects in a multiresolution representation. The paper concludes by describing a complete system that could integrate all these concepts, and over the years many Sarnoff products have included implementations of these ideas (e.g. the pyramid Vision Machine, the Acadia processor, etc.). Burt [61] reviews a suite of pyramid-based algorithms and a progression of vision processors developed based on them. Conception and Wechsler [62] present an implementation of many of the key ideas of Burt [60], with a coarse segmentation and classification stage followed by a memory-driven multiresolution recognition stage reminiscent of Selective Tuning [63], but applied on a wavelet pyramid.

Howarth and Buxton [64] investigate the use of Bayesian inference networks to provide a task based control system identifying relevant objects in the scene which potentially fulfill the given task. The system accepts sequences of images representing traffic movement, and the task is to identify areas of interest and determine relationships between these areas (e.g. overtaking, following, etc.). The computational load on the inference system is reduced since it is not necessary to consider the interactions between all objects in the scene.

Baluja and Pomerleau [65] trained a neural network to perform dynamic relevance assessments by using the hidden layer representation to predict the next input image, in a manner very similar to Kalman filtering. The system is used in three very different tasks. The first task is autonomous driving, the system being used to monitor lane markings. The attentional focus mechanism allows the network to work in real lighting conditions, with typical street markings that have frequent misleading features. The second task, anomaly detection in silicon wafers, demonstrates the system's ability to highlight unexpected features in the input image. The third application, hand tracking in cluttered environments, demonstrates the fact that the system's design allows a priori task information to be integrated into solutions.

Sela and Levine [66] and Gallet et al. [67] present approaches based on the detection of symmetries in input images. In [66] the focus is on a novel real-time algorithm for computing points of interest defined as the points of intersection of lines of symmetry between edges in the image. The main advantage of the algorithm seems to be its ability to select a relatively small number of points of interest, both when dealing with faces and when dealing with complex outdoor scenes, points of interest that, at least subjectively, seem to correspond to relevant areas of the image. By contrast, the algorithm presented by Gallet et al. selects points of interest by pairing oriented curvature points [67]. Since this method generates many points of interest, local competition is employed to select the most relevant ones. This algorithm has been used successfully to guide vision-based robots through an office environment augmented with directional markers [68].

Real-time motion tracking is obtained in Toyama and Hager by combining a series of visual search and tracking algorithms in a hierarchical system [69]. The attentional mechanism monitors performance and performs two tasks: rapid selection of potential candidate locations and the selection of the most appropriate algorithm to execute. Both theoretical and empirical evaluations of the method are provided.

While the preceding systems use low or intermediate level features in determining the points of interest within the image (or image sequence), Sun and Fisher [70] introduce high-level features and perceptual grouping into the analysis. Competition occurs at all levels, from features within objects to groups of objects, and selection can similarly occur at all levels. Applying external biases, the system can perform tasks as varied as region segmentation and object recognition.

This brief analysis of what is a relatively small subfield highlights the potential and broad applicability of attentional methods in computer vision.

One final point to mention is that the use of attention is not limited to computer vision, other areas of computer science benefit from the idea of selective processing to reduce computational load. For example, an attention-based communications scheme named “progressive transmission” has been proposed by Zabrodsky and Peleg [71]. Under this proposal, data is first transmitted at low-resolution, and then it is up to the recipient to request any additional high-resolution data based on task requirements. NASA uses a simplified version of this approach in its communications with space probes.

1.3.5. Conclusions

Teaching computers to see is a very difficult problem, and computationally very expensive. It is thus not surprising that despite the wide diversity of the field, one common theme seems to run through most research: the reduction of the computational load through selective processing. A few main themes have emerged. Active vision attempts to maximize the relevance of the images by proper placement of the cameras, in order to simplify the task of the modules of the system. Perceptual grouping attempts to move away from pixel-level information, thus allowing the system to work on significantly fewer basic features. The selective processing of areas of interest is probably the most similar approach to visual attention, and some results are very impressive, but the fact that semantic information is generally not used because of the very low level features used limits the applicability of the method.

1.4. Key experimental results

Any biological modeling effort needs to start from experimental data, and visual attention is no exception. A wealth of information has been published, using a wide variety of methods, to the point where a thorough review is beyond the scope of any one publication, but a short sampling of important experimental paradigms and results will serve to highlight some of the key concepts and important features that any theory and model must address, while at the same time touching on some of the controversies in the field and some of the many open questions that need to be addressed.

1.4.1. Visual search

Probably the simplest method used to investigate visual attention involves visual search tasks, and the evolution of many of the visual attention theories can be correlated to seminal results of such experiments. In the standard visual search experiment, subjects look for a target item among a number of distracter items (the total number of items in the display is named set size). Generally the target is present in half of the trials while in the others, only distracters are presented, and subjects have to indicate whether or not the target is present. As the number of distracters increases, reaction times typically also increase, and the rate of this increase is of particular interest.

Early experiments produced a very nice picture of a dichotomy between “parallel” and “serial” searches. In “parallel” searches, reaction times are independent of the number of distracters, and the usual interpretation of this result is that all items are processed at once to a level that is sufficient to distinguish targets from distracters. In “serial” searches, reaction times increase with the addition of distracters at a rate of approximately 30 ms/item for target present and about 60 ms/item for target absent trials, a pattern that is consistent with a self-terminating search through the items.

Later experiments have shattered this simple image, with experimental results showing a continuum of reaction time vs. set size slopes, leading Wolfe to propose the description of search performance in terms of efficiency classes, from “efficient” searches characterized by a slope of approximately 0 ms/item, through “quite efficient” at 5-10 ms/item and “inefficient” at 20-30 ms/item, to “very inefficient” searches characterized by slopes of more than 30 ms/item [72].

A very influential series of experiments has been proposed by Quinlan and Humphreys [73]. They define different kinds of search tasks in terms of pairs of numbers of the form (m,n), where m is the number of features and n is the number of features by which each distracter group differs from the target. In this framework, (1,1) corresponds to feature searches, (2,1) to standard conjunction searches, (3,1) and (3,2) are triple conjunction searches where the target differs from all distracter groups by one and two feature dimensions respectively. Many computational models of visual attention have used this type of experiment to evaluate their performance.

Issues are complicated further by the existence of search asymmetries, which occur when a target item of type A among distracters of type B is easier to find than a target of type B among distracters of type A. For a thorough review of search asymmetries see the special issue of Perception and Psychophysics (vol. 63, issue 3, 2001), and also note that Rosenholtz argues that in most cases search asymmetries are just an artifact of hidden asymmetries in the experimental design [74].

For thorough reviews of the visual search literature, see [72, 75].

1.4.2. Saliency

Two aspects of saliency have been described: global saliency maps and local saliency modulation. Saliency maps, initially proposed by Treisman and Gelade in the Feature Integration Theory [76] as "master maps of locations," and, in the current interpretation, by Koch and Ullman [77], have been hypothesized as solutions to the problem of merging the multiple feature-specific early representations into a single attentional focus. Local saliency modulation has been observed in the striate cortex and is interpreted as either a precursor or a consequence of texture and/or object segregation.

A number of anatomical sites have been proposed as possible locations for a global saliency map, including the superior colliculus, the lateral geniculate nucleus, the posterior parietal cortex, the dorsomedial region of the pulvinar, and the primary visual cortex. The strongest candidate seems to be the lateral intraparietal (LIP) area [78]. This area is known to be important in attentional and oculomotor processes, and it has connections with the frontal eye fields, superior colliculus (both areas involved in the generation of saccadic eye movements), and to prestriate and inferior temporal visual areas. Neurons in the LIP of the rhesus monkey have been found to respond to recently flashed stimuli better than they respond to stable, behaviorally irrelevant stimuli. They respond transiently to abrupt motion onsets, but have no directional selectivity. The conclusion of the study is that LIP is important in the attentional mechanisms preceding the choice of saccade target rather than in the intention to generate the saccade itself, and thus they seem to represent salient stimuli, but whether this is sufficient to declare LIP "the" saliency map is subject to significant debate.

The recent study presented in [79] (on monkeys of an unspecified species) draws similar conclusions, and also notes that the locus of attention cannot be ascertained by measuring the activity of a single neuron in LIP, or even by measuring the activity of all the neurons whose receptive field overlap on a location or object. A global analysis is needed of the LIP neurons that represent the whole visual space, in which case activation at one site can be correlated to the saliency of the associated image area.

Physiological and psychophysical studies show modulation of responses of primary visual cortex neurons that is consistent with contour enhancement, figure-ground segregation, and texture segmentation. Recordings from monkey V1 analyzed in [80] show that collinear segments outside the classical receptive field of a neuron enhanced its response to the preferred stimulus, and the same effect was shown in contrast thresholds in humans. Gallant et al. summarize research results on V1 responses to textures, observing that responses along texture borders are enhanced, consistent with texture segmentation, and that neurons corresponding to the “inside” of a texture defined area are also enhanced, but to a lesser degree, consistent with object-background segmentation [81]. See [82] for additional details.

While these results are very well established, there is significant disagreement in the literature about the mechanisms responsible for them. Some authors propose the lateral connections within V1 [83, 84], while others suggest that at least texture segmentation and figure-ground segregation occur in higher cortical areas (V2, V4) and the modulation observed in V1 is due to feedback connections [85, 86].

1.4.3. Neurophysiological mechanisms

A great deal of research has focused on the neurophysiological mechanisms of visual attention.

Sensitivity to stimuli at attended locations is increased, and, at the neuronal level, this could be explained by a multiplicative increase in firing rate or by an increase in the effective strength of the stimulus. Each explanation results in different and conflicting predictions. To test these predictions, Reynolds et al. recorded responses of macaque V4 neurons to stimuli across a range of luminance contrasts and measured the change in response when monkeys attended to them [87]. It was observed that attention caused greater increases in response at low contrast than at high contrast, which is consistent with the predictions of the increase in effective stimulus strength model. This effectively means that at the level of individual neurons, the effects of attention and luminance contrast are indistinguishable. The same conclusion was reached in two independent studies on attention to motion in macaques [88] and in humans [89].

McAdams and Maunsell examine the influence of spatial attention on the orientation tuning of neurons in areas V1 and V4 of rhesus monkeys [90]. Attention was found to enhance the responses of neurons in both areas, but the width of their orientation-tuning curve was not systematically affected. Thus, the effects of attention at the individual neuron level are consistent with a multiplicative scaling of the response across orientations. Similar results are reported in area MT for non-spatial, feature-based attention [91].

1.4.4. Neurophysiology of attentional competition

Given the hierarchical nature of the visual cortex, characterized by progressively larger receptive fields, one of the most important questions that needs to be answered is what happens when two stimuli fall within the receptive field of one neuron. In the classical study of Moran and Desimone, monkeys were trained to attend to stimuli at one location and ignore stimuli at another [92]. Single cell recordings in areas V4 and IT revealed that when both stimuli were within the receptive field of the same neuron, the effect of the unattended stimulus on the neuron’s response was significantly reduced.

Since then, it was repeatedly shown that the greatest attentional modulation is observed when multiple stimuli appear within a cell’s receptive field. To quantify this effect, Reynolds et al. measured the responses of neurons in macaque areas V2 and V4 to various combinations of stimuli, with or without attention [93]. As a control, they measured each cell's response to a single stimulus presented alone inside the receptive field or paired with a second receptive field stimulus, while the monkey attended to a location outside the receptive field. The results confirmed the earlier finding that adding the second stimulus causes the neuron's response to decrease. Directing the monkey's attention to one element of the pair, the neuron's response moved toward the response elicited when the attended stimulus appeared alone. The authors see this result as consistent with the idea that attention biases competitive interactions among neurons, causing them to respond primarily to the attended stimulus.

1.4.5. Overt and covert attention

Many experimental setups require the subject to fixate and avoid eye movements in order to eliminate confounding factors or to exploit the retinotopy of many areas in the visual cortex. This is a very unnatural situation, normally primates make 3-5 saccades per second in order to foveate areas of interest in the environment. A wide variety of studies indicates that the two forms of attention share many of the neural structures involved, including areas in the superior colliculus, the pulvinar, the frontal eye field, the precentral gyrus, and the intraparietal sulcus in the macaque, and homologues and/or adjacent areas in humans.

In the area of psychophysics, [94] investigates the relationship between saccadic eye movements and covert orienting or visual spatial attention. Subjects were required to make saccades while attending to a target or location, and the results show that subjects cannot move their eyes to one location and attend to a different one. In another experiment, subjects were instructed to expect a cue at one of four predetermined locations, thus directing covert attention to that location, and make a saccade down when the cue appears [95]. The trajectory of the saccades deviated contralateral to the hemifield in which the imperative stimulus was presented, demonstrating that spatial attention allocation leads to an activation of occulomotor circuits, in spite of eye immobility. Kowler et al. have found that saccades are facilitated by covert attention, perceptual identification is better at saccadic goals, and attempts to dissociate the locus of attention from the saccadic goal were unsuccessful, i.e. it was not possible to prepare to look quickly and accurately at one target while at the same time making highly accurate perceptual judgments about targets elsewhere [96]. The results are explained by a model in which perceptual attention determines the endpoint of the saccade, while a separate trigger signal initiates the saccade in response to transient changes in the attentional locus.

Physiological studies show similar results. Colby and Goldberg found that while some parietal neurons represent object locations in motor coordinates, and the salience of a stimulus is the primary factor in determining the neural response to it, visual responses are independent of the intention to perform saccades [97]. Andersen et al. study the effect of eye position on the light-sensitive, memory, and saccade-related activities of neurons of the lateral intraparietal area and area 7a in the posterior parietal cortex of rhesus monkeys [98]. The study finds that activity depends on the vector from the current eye position to the cue or movement end point location, again independent of the intention to perform saccades.

While this sharing of resources is fairly well established, overt shifts of visual attention pose the additional problem of the need to remap the internal representation of the world, and in particular, of the saliency map, and little is known with certainty in this area.

1.4.6. Top-down influences

When viewing a natural scene or a visual stimulus, a number of factors influence the deployment of attention. In general, salient items, or items that are different from their neighbors, tend to attract attention, and the information that guides attention in this case is purely stimulus-related. It is hard not to notice and attend to a red stimulus embedded in a field of green distracters. This type of information is termed “bottom-up.” On the other hand, we don’t have any difficulty in searching for particular orientations amongst the green distracters in the same stimulus. The stimulus did not change, but our intention did, and the type of information that guides attention in this case is labeled “top-down.” Several forms of top-down information are identified in the literature. Explicit information, as presented above, can take the form of verbal instructions or image cues. Spatial information can guide attention to specific locations within an image, while implicit information is based on previous stimuli. The latter can take the form of “priming of pop-out” [99] when subjects respond faster to a feature if recent targets have been the same feature, or “contextual cueing” [100] 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 [75], here we will only look at two dramatic illustrations of the importance of top-down attention on perception.

In a study aimed at measuring the impact of top-down attentional selection, Blaser et al. used an ambiguous illusory motion stimulus whose motion can be influenced by attention [101]. The stimulus consists of a temporal sequence of five frames, each containing a vertical sinusoidal grating. The stimuli used two types of grating, a red-green isoluminant grating and a contrast-modulated random noise grating. The motion sequence was constructed by alternating red-green and noise frames, with each frame displaced 90° relative to its predecessor. The background was a 50/50 mixture of red and green (i.e. yellow). No attentional instructions were given in the control experiment, and psychometric curves of perceived motion direction detection were generated by modulating the chromacity difference between the color patches and the background. After this, subjects were instructed to repeat the experiment while attending to one of the colors, and it was observed that the psychometric curves were shifted in the direction of motion consistent with the attended color. The authors interpret the lateral shift of the curves as indicative of the magnitude of the attentional effect, and note that this magnitude does not depend on the spatial frequency of the stimuli.

While most neurophysiological studies focus on neuronal activity in the intervals following the onset of the trial, one of the most interesting findings of [102] refers to cell activity preceding the presentation of the cue, so before the actual start of the trial. In their experiment, rhesus monkeys were trained to fixate a cross in the center of a screen. Trials started with the presentation of a complex cue (one of 24 complex color images, ranging from identifiable objects to colored textures and patterns), followed, after a brief delay, by an array of images from the same database. The monkeys were trained to respond to the identification of the cue within the array either by performing a saccade to the image, or by releasing a lever. Two stimuli were selected for each neuron based on the response elicited in a simple fixation task: a "good" stimulus, which evoked a strong response and a "poor" stimulus, which evoked little or no response. As in other studies, neuronal responses following a "good" cue were seen to be elevated in the delay period, indicating a memory for the task stimulus. Interestingly, when comparing responses preceding the presentation of the cue in random and blocked design sets of trials, it was observed that activity in the latter case was significantly above baseline when the monkey had reason to expect the "good" cue. To eliminate the possibility that this increase is an artifact of nonspecific changes in cell activity across the session, blocks of trials with the "good" and "poor" cue were interleaved. Also, the higher activity preceding the "good" cue in the block design could not have been due to a lingering response to the "good" stimulus as target on the previous trial since this effect was not observed in the random design. The authors concluded that the sustained activity preceding the cue in the blocked design was a purely "cognitive" phenomenon related to expectation of a specific cue and could not be a sensory response.

1.4.7. Objects and attention

In most traditional theories and models of visual attention, attention is characterized in spatial terms as a "spotlight" [103], "zoom lens" [104] or "beam" [63], and many studies use spatial cues to direct attention to a particular spatial location. It is well established that valid location cues speed the response to a target while invalid cues slow it down [103, 105]. But even early on, studies into what was called "selective looking" provided evidence that even when stimuli are spatially superimposed, attention can select one of the stimuli and ignore the other [106-108]. Some recent studies have confirmed and expanded on this result [109, 110], while others have used the effect to test the validity of theories and propose new ones [91].

One of the most influential experimental paradigms in the study of object-based attention was proposed by Egly et al. [111]. In a typical experiment, the display consists of two bars, and subjects were cued to one end of one of the bars and the task was to detect changes at another location, on the same or on the other bar, at equidistant points. In what was termed "same-object advantage, " subjects consistently respond faster to changes on the cued object, even though the spatial distance between the cued location and the two probe locations is the same. Even more convincing, [112, 113] have shown that the effect survives occlusions.

Other evidence for object-based attention comes from a number of disorders such as unilateral neglect [114, 115] and the Balint syndrome [116]. Unilateral neglect patients typically have lateralized parietal lesions and fail to perceive stimuli in the visual field contralateral to the lesion. While the effect seems to be mainly spatial, two types of experiments suggest that it could also be object-based. Caramazza and Hillis have shown that some patients neglect the contralesional half of object with salient axes regardless of the visual field in which they are presented [117]. Behrmann and Tipper have used dumbbell type objects. Initially the patients were slower to detect targets on the contralesional disk, but after the object was rotated in front of them by 180°, the same patients responded faster to the now contralesional disk [118, 119]. Control experiments have shown that the effect is indeed object-based: after removing the connecting line the effect disappeared, and the effect did not transfer to stimuli added to the background. Some patients with Balint's syndrome exhibit a condition known as "simultanagnosia," which is the inability to perceive more than one object at a time, despite normal visual processing, even when the objects are spatially overlapped. If two overlapping triangles forming a Star of David are colored differently, patients often perceive only one of them.

For additional information on object-based attention see [120].

1.4.8. Conclusions

It is obvious from this brief and selective review that visual attention research is a very broad area, and that the phenomenon manifests itself at all levels of investigation. As such, the task of the modeler is very difficult. The breadth of research and the sometimes contradictory results make it highly unlikely that a single model can account for all observed phenomena. Models concentrating on the overall functionality revealed by psychophysics can not shed light on the neural mechanisms involved. The sheer volume of the primate visual system and its tight integration with the rest of the brain make it very difficult for detailed models based on neurophysiology to scale up to the level where they can meaningfully simulate behaviour.

2. Theories of Visual Attention

In this section we will review some of the major theories of attention. This brief overview is not intended to be exhaustive, it merely attempts to identify what we see as the three major philosophical approaches to attention, and provide a theoretical context for the review of computational models. For a thorough historical review that tracks the evolution of the major ideas of the field, see [121].

Even a summary review of the relevant literature shows that visual attention is a very broad and fragmented field of study. Because of this, and to provide some context for the rest of the paper, we will start by listing what we see as the key questions that need to be answered by any work that wants to claim to be a theory of attention. Since, as far as we know, no theory or model comes anywhere near meeting these criteria, and borrowing from physics, we will call this a "grand unified theory of attention" – similar in a way to Newell's unified theories of cognition [122].

2.1. Key ingredients

To account for the wealth of experimental data in a single theory is not an easy task, and as such, it is probably not surprising that most theories of attention focus on explaining particular aspects of this complex phenomenon. To be in a better position to compare the various theories and models and trace their evolution, we will start by listing the components that seem to be needed for a grand unified theory of attention.

A theory of attention must be able to explain salience and pop-out. While some authors don't consider this to be a component of attention per se, but a mere side effect of preprocessing steps or of the intrinsic connectivity patterns, the mechanisms involved are very likely either shared or at the very least intimately related. Some theories and models, such as biased competition, actually claim that the same mechanism is responsible for both salience and attentional selection, considering attention an emergent property of the competitive dynamics of the system.

A theory of attention must also be able to explain the mechanisms of 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 representation and how they can be selected as the subject of attention.

Theories of visual attention must define the kinds of stimuli that can be attended. This is a very broad question, touching on issues such as attending to locations vs. objects vs. features, the size of the attentional field, the shape and contiguity of the "focus" of attention – can attention select discontinuous regions? Can it select doughnut shaped regions? Can attention select multiple locations or objects? What are the mechanisms responsible for these properties?

Another important component of a unified theory of attention is attentional selection, i.e. what exactly does it mean that a particular stimulus is selected by attention? This is the interface between perception and consciousness, and relatively little is known with any degree of certainty. 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. 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 and reach consciousness, but in general, the information that gets through is either incomplete, or incorrectly bound. 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 the focus of attention, an important question is what effect does this have on future selections? As with any other topic in visual attention, opinions are divided. The classical view requires a mechanism of inhibition of return [123] while a number of controversial studies seem to suggest that memory plays only a marginal role in visual search [124]. While some of the evidence presented is indeed tantalizing, the initial research has been fraught by methodological mistakes and exaggerated claims that don't help in making the idea very popular.

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, seems to dominate. Various studies seem to indicate that the two are closely related and that the same neural structures involved. At least two aspects of this dichotomy need to be addressed: what are the similarities and differences between overt and covert shifts of attention? And, in the case of overt shifts of attention, what are the mechanisms that compensate for the shifts in the saliency map that accompany eye movements?

Last but not least, what are the implications of attentional selection from an object recognition point of view. It seem that attention is needed to recognize certain categories of stimuli and not others, 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.

Of course, this fine grained analysis of the field is important in judging the completeness of the various theories, and it can provide a framework for generating and interpreting predictions, but it is not very useful in getting a clear picture of the fundamental underlying assumptions made by each theory. We have identified three major philosophical approaches and in the next sections will present the main theories of visual attention using this classification.

2.2. Attention as selection

Many researchers view attention as a selection mechanism, used to simply identify areas of interest in the visual scene and make them available for further processing. This could take the form of object recognition, memory, or conscious perception. Most theories of visual attention fall under this category.

2.2.1. Early vs. late selection

Early work in attention was based mainly on auditory perception, a fact reflected in the terminology used and in the format of the theories that were advanced.

The first major theory of attention was Broadbent's Filter (or early selection) theory [125]. Its basic hypothesis is that the recognition mechanism is only able to handle one stimulus at a time, and thus, after an initial representation of all the physical attributes of the input, a selection criterion is used to determine which stimuli should be processed further. Because the attentional selection is supposed to act early (i.e. before stimulus recognition), this and other similar theories have been termed early selection theories.

At the other extreme, [126-128] have proposed that recognition has no capacity limitations, but occurs in parallel for all stimuli, and selection is only needed for access to memory and consciousness.

The idea that selective attention is more flexible than either the early or late selection models allow, first introduced by Treisman [129], was detailed in the multimode model [130]. Based on experimental evidence presented by Johnston and Heinz [131], they proposed that attention could operate at various stages of processing (or, as they put it, in multiple modes – early, middle, late) and that selection will occur as early as possible depending on the task demands. This is due to the hypothesized higher cost of later selection, which makes it less accurate than middle or early selection. Due to its lasting impact on the field, and especially in the computational modeling world, dominated by the Koch and Itti early selection style model, and the fact that later research has shown it to be incorrect, this early serial/parallel dichotomy has been deemed to be a "useful, but potentially dangerous fiction" [72].

2.2.2. Feature Integration Theory

Feature Integration has evolved significantly from its earliest incarnation - the Attenuation Theory, which is a typical example of a late selection theory. The key points in the theory's evolution are the idea that attention allocation is gradual rather than binary, attenuating unattended signals [126], followed by the idea that attention is a hierarchical process, operating at various level [129].

As the field started moving from auditory to visual attention, this set of ideas matured and incorporated new experimental results. In the Feature Integration Theory [76] we see some of the major themes that have dominated most of the field ever since. The model is hierarchical, with early processing occurring in separate modules for each feature dimension, and consisting of the detection and comparison of features. The selection is done by an attentional window that can operate at a specific stage. Feature Integration Theory interprets the dichotomy between "serial" and "parallel" search indicated by the early results quite literally, the model keeping a very clear distinction between preattentive and attentive mechanisms. The preattentive mechanisms are supposed to operate in parallel on the whole image, and are responsible for flat search times, while the limited capacity attentive stage is responsible for the serial search.

While some of the explanations provided have been revised, the concepts of saliency, attentional spotlight, pop-out and serial search proposed by Feature Integration Theory continue to be enduring themes of attention research.

2.2.3. Guided Search

Guided Search [132-134] maintains the preattentive-attentive separation, but removes the need for an explicit distinction between parallel and serial search. In the first stage of processing, features are segregated in parallel and engage in a limited form of competition with their neighbors. The second stage is a weighted sum of the feature maps that builds a saliency map, and this map "guides" the deployment of the limited attentional resources. To explain primate performance in visual search tasks without the need for distinct parallel and serial mechanisms, Guided Search postulates that the early stage processing is noisy, which makes the guiding process less precise. Top-down guidance is in the form of biasing the weighted sum towards the desired features.

2.3. Attention as filtering

One of the characteristics of a hierarchical system that does image processing by introducing scale and location invariance, as the primate visual system seems to operate, is that neurons in the higher levels of the hierarchy have large receptive fields that are likely to include not only the optimal stimulus but also significant amounts of extra information. The proposed role of attention is to filter out this information in order to improve the signal-to-noise ratio in the system.

2.3.1. Selective Tuning

While most theories are more or less attempts at explaining experimental results, the Selective Tuning theory [63] is unique in that it is based on a rigorous first principles analysis of the theoretical aspects of biological image processing, and in particular on a complexity analysis of the tasks involved [2]. Selective Tuning is based on a hierarchical processing pyramid with reciprocal connections that carry stimulus information forward and attentional control signals backward. According to the theory, the role of attention is to aid stimulus recognition by reducing the interference caused by the large receptive fields found in higher visual areas. In order to do this, stimuli proceed to the higher levels unimpeded in a first pass, and as a series of winner-take-all computations are propagated backwards, any items that could potentially interfere with high level representations are actively inhibited. The attentional control is both local and distributed.

Probably the most powerful feature of Selective Tuning is the theory's ability to generate predictions that have received significant support:

- An early prediction ([2]) was that attention seems necessary at any level of processing where a many-to-one mapping of neurons is 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 [92]. Since then, attentional modulation has been found in many other areas both earlier and later in the visual processing stream [135]. Evidence cited by Britten [136] who reached the conclusion that 'attention is everywhere', was mostly post-1990. Vanduffel et al. [137] have shown that attentional modulation appears as early as the LGN.
- Another early prediction of Selective Tuning is that attentional modulation in higher areas precedes that in earlier areas, a prediction supported (at least in the ventral pathway) by Mehta et al. [138].
- The notions of competition and of attentional inhibition were also early components of the model [2] and this too has gained support over the years [93, 135, 139].
- The model has always included an inhibitory surround component [2]. This implies that perception may be negatively affected in the vicinity of the attended stimulus. This too has recently gained support [137, 140-142].
- The model also explains how so-called pre-attentive vision is only a special case of attentive processes [2]; no separate pre-attentive process operates independently of attention, a view Joseph et al. [143] seem to be suggesting too.

2.4. Attention as process

A number of theories discard the notion of a special attentional mechanism and attribute all of the observed characteristics to emergent properties of the competitive neural networks present in the brain.

2.4.1. Biased Competition

While all the theories reviewed above maintain in one way or another the notions of saliency and attentional window, Biased Competition [139] introduces the idea that attention is an emergent property of the dynamics of the visual system. Their theory introduces the notion that all the stimuli in the visual field compete for access to computational resources to influence behaviour. This competition takes place at all levels, and it can be influenced by top-down biases both at a spatial and at a feature level. The result is that the representation of behaviorally irrelevant stimuli is suppressed.

Biased Competition accounts for various slopes of search times by appealing to the notion of similarity between target and distracters and the similarity of the distracters to each other. Unlike selection theories that treat features and conjunctions differently, resulting in a vast literature that attempts to list all the basic features, Biased Competition argues that they are qualitatively the same.

One of the most controversial proposals of this theory is the notion of receptive field plasticity, attention in effect "shrinking" the receptive fields around the attended stimulus.

2.4.2. Feature similarity gain

An alternative to biased competition has been proposed by Treue and Martinez-Trujillo [88, 91], based on electrophysiological data obtained from areas MT and MST of macaque monkeys, data that was in disagreement with predictions of the biased competition theory. Their studies address both the issue of spatial attention and feature based attention, and suggest that spatial and feature based attention modulate neuronal responses without changing the cell's selectivity, and that this modulation changes a given neuron's response depending on the similarity between the attended feature and the cell's preferred feature.

The main experimental result is that attention only modulates the gain of neural responses via a multiplicative mechanism that does not affect the shape of the tuning curve of individual neurons. This means that instead of changing or biasing the receptive field, attention merely changes the salience of the physical stimulus, so it is equivalent to and indistinguishable from an increase in the intensity of the stimulus. Experimental support for this ranges from neurophysiology to psychophysics [89], and [88, 144] propose and model a physiological mechanism that could potentially account for this effect.

To address the issue of receptive field plasticity, Treue and Martinez Trujillo used superimposed random dot patterns moving in different directions, and the monkeys used in the experiment were trained to selectively attend to one of the directions [88, 91]. The modulatory effects did not change from a single stimulus scenario, and as the random dot patterns were spatially superimposed, receptive field plasticity cannot account for this. In addition, once animals attended to a particular stimulus feature, attentional gains for that particular feature were found throughout the visual field, in contradiction with the notion of receptive field modulation.

2.5. Conclusions

In this chapter we have introduced an original classification of the major theories of visual attention, corresponding to a certain extent to the historical development of neuroscience.

Selection theories present a very attractive, simple and intuitive picture. The notion of a localized saliency map in the brain is very compatible with the modular and reductionist approach that has characterized much of neuroscience in the past. These theories seem to be able to account for some behavioural results, especially in visual search, but they offer little, if any, insight into the neurophysiological aspects of attention in the main processing streams of the visual system.

Filtering theories focus on the process of extracting meaningful information out of cluttered scenes. A number of facts converge on the conclusion that Selective Tuning has the potential to make significant contributions to both theories of visual attention and computer vision. Selective Tuning is the only theory based on rigorous formal analysis (namely computational complexity theory). This approach moves attention in the direction of signal processing and information theory, and is in a unique position to take advantage of results in these areas. Also, the level of the theory is such that it can bridge the gap between neurophysiology and psychophysics.

The emergent property theories share some of the advantages of filtering theories, but lack the formal underpinning, and seem to be developed as reactions to developments in the field, rather than being active generators of predictions. Decentralized processing seems to be a major direction in current science, and developments in this area could provide these theories with the necessary formal support.

3. Computational Models of Visual Attention

3.1. Classes of computational models or What is included, what is not?

“Modeling plays a unique role in visual neuroscience. On one hand, single-unit physiology catalogues neural responses to visual stimuli and perhaps correlates them with behavior; on the other hand, psychophysics measures the overall functional capabilities of the visual system. The goal of neural modeling is to produce theories that bridge the gap between activity of the neural elements and responses of the functioning system. Thus, an ideal neural model contains elements mimicking the response properties of neurons in relevant visual areas and makes predictions of system behavior that are psychophysically testable“ [85].

Wide varieties of computational models of visual attention have been proposed, including purely mathematical (such as [145]), signal detection (such as [146, 147]) and Bayesian models (such as [148]). Some of these models provide very good fits for the experimental data and are able to generate interesting predictions, but they fail to provide the bridge required by Wilson’s definition, and are not providing explanations for the neural mechanisms involved. Only models that fit this definition will be included in this review, because in the author’s opinion, these are the models that are most likely to generate significant progress in neuroscience.

3.2. Computational models

In the remainder of this review, instead of providing a phonebook style list of the current biologically plausible computational models of visual attention, we will try to see how they address the key ingredients of an attentional theory identified above.

3.2.1. Saliency maps

Most computational models of visual attention use saliency maps 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.

In the Shifter Circuits model [149, 150], inputs are filtered through Gaussian blurring creating a "blobs map” in which units compete with each other and only the control unit corresponding to the strongest blob prevails. In what is a clear example of an “attention as selection” model, the input corresponding to the strongest blob is selected and routed to an object recognition module.

The visual attention model of Itti and Koch [151] is inspired by the local center-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.

One of the problems encountered by computational models that use saliency maps is that they translate physical properties of the stimulus such as luminance, color, size, onset, etc. into saliency values. Because 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 [152]: simple normalized summation, linear combination with learned weights, global non-linear normalization followed by summation, and local non-linear competition between salient locations. The conclusion of this study is 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 and Lionelle’s critical evaluation of the [151] model starts from the assumption (questionable in the author’s opinion) that attention is simply a front-end for object recognition [153]. If this assumption is true, Draper and Lionelle 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. Draper and Lionelle’s 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.

An example of linear combination with learned weights is [154], in which a backpropagation network receives task-dependent input and controls the flow of information from the low level feature maps to the saliency (“priority”) map.

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. with very impressive results, but at the expense of generality [155]. Their Interactive Spiking Neural Network (ISNN) is geared specifically to 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-color 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.

Similarly, in order to facilitate segregation and object-based attention, [156] uses symmetry, eccentricity, color contrast and depth to construct a 3D saliency map.

Interestingly, while neurophysiologycal evidence for saliency maps in the brain has been mounting (e.g. [157]), a number of models have started arguing that they might not be needed after all, and proposed attention as an emergent property of the dynamics of the object recognition network.

In the solution presented in the Selective Tuning Model [63], 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] (Θ-WTA) competitions. These competitions refine the very coarse initial representation of the attended stimulus location, 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 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 [158] 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, and the dynamics of the network produces winners without the need for explicit representation of salience. This model is split along the ventral/dorsal divide, with the ventral 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 modeling 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.

The underlying neural mechanisms of this competition are investigated in [93], starting from the observation that attentional modulation peaks when multiple stimuli share a neuron's receptive field. The authors propose a very simple model neural circuit, composed of three neurons, two input and one output, connected both directly and through inhibitory interneurons. The input neurons correspond to the reference and probe stimuli, and the dynamics of the circuit is described through differential equations. Attention is assumed to increase the strength of the signal coming from the cell activated by the attended stimulus, through an unspecified mechanism. The circuit behaves like a feedforward competitive network, and is able to simulate the observed behaviour of its biological counterpart. Work on the computational modeling of the competition and selection microcircuitry has also been reported by Grossberg [83, 159] and Li [84, 160].

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

3.2.2. 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.

Designed specifically as a word recognition system, MORSEL [161] integrates top-down influences in a task-specific fashion. As discussed in more detail in the "Attention and recognition" section, the system employs two distinct attentional selection mechanisms: a late selection component (a "pull-out net") and an early selection component (the "attentional mechanism"). The top-down component of the attentional mechanism moves the spotlight based on specific information such as static target expectations or dynamic scanning patterns for reading. A second type of top-down influence, independent of attention, can suppress features related to distracters in search tasks, thus reproducing the flat search times observed in psychophysics experiments.

As seen above, in the Selective Tuning Model [63], 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 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 scanpaths, i.e. a qualitative evaluation of the system's performance, are presented, and not reaction times in search tasks. Recent work [162] has started to address this issue by explicitly trying to model Motter's monkey experiments [163] using cameras mounted on a robotic head.

A model aimed specifically at the aspect of learning top-down influences uses a backpropagation network that receives task-dependent input and controls the flow of information from the low-level feature maps to the saliency (or "priority") map [154]. The network is trained to enhance the relevant and suppress the irrelevant information for the current task. As shown by [152], this approach can have good results in dealing with specific problems at the expense of generality.

In the Neurodynamical Model of visual attention [158], 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 modelers 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 [164], which is a spinoff from previous work by the same researchers on rapid scene categorization (e.g. [165]), which in turn follows the pioneering work of Biederman [166]. The key point of this previous research has been to demonstrate that low-resolution 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 whether this type of cueing constitutes "top-down" influences, the approach has significant merit, as it and [158] 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 [155], 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 analyzed formally by [167].

3.2.3. 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 [151] 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 [63, 168, 169], 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 soft winner-take-all competitions with proximity biases.

In general, it is difficult for modelers 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 the theories and models, with the notable exception of the Selective Tuning Model [63]. The fundamental theoretical assumption behind this model 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 at all 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. As mentioned before, this prediction of the model has recently received significant experimental support [142, 170, 171].

Attentional control has not been the focus of intense modeling research, if it is mentioned at all, it is in the form of unspecified external mechanisms. For example, in the Neocognitron system [172] external switch signals disengage attention and allow it to focus on the next target. Similarly, in the Selective Tuning Model [63] the time course of the attentional process is determined by "gating control signals" of unspecified origin. These binary signals are responsible for initiating the WTA processes in the appropriate sequence and for determining the duration of one attentional fixation.

3.2.4. 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 [123].

In terms of IOR in covert attention, three types of solutions have been presented in the literature. Some models just inhibit selected locations, either in the input image [63][2] or in the saliency map [151]. 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 [172], is to simulate neural fatigue, controlled by external attention switch signals. A third solution, presented in the Selective Tuning Model, disables the neural pathways corresponding to the selected item or location [63].

Models that use distributed representations of salience in the form of a dynamic winner-take-all (WTA) network might not even need an explicit IOR mechanism. This issue is not discussed in the computational modeling of visual attention literature, but models of the dynamics of small neural networks prove that integrating neural adaptation in WTA networks produces exactly the type of short-term memory that seems to be needed for IOR [173].

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 and some form of short-term spatial memory. In an explicit attempt to address the issue of moving objects, [156] implements a "semi-attentive" stage, which is in effect 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, unselected objects having the highest priority. While the idea of object files might have some biological support, these object files are implemented as symbolic representations of position, size, trajectory, etc.

At the same time, without specific reference to attention, a number of models have addressed the issues of coordinate transforms and/or dynamic remapping that seem to be needed (see [174] for a review).

[2] Implementation only, the theory uses the third solution.

Two recent publications showcase the state of the art in this area. Zaharescu et al. investigates the neural mechanisms for saccade target determination and execution, proposing that processing in the central part of the field of view provides object recognition, while coarser processing in the periphery provides support for saccadic eye movements [162]. While the decisional aspects of the proposed model are rather simplistic, it is the first model to integrate a full spectrum of mechanisms, including coordinate transforms, saccadic remapping, and inhibition of return. Complementary research is presented by Lanyon and Denham, where the decision-making surrounding the execution of saccades is investigated in detail. Starting from feature-based attention, the model proposes a pathway that involves LIP, where behaviorally relevant locations are represented and saccadic eye movement decisions are taken [175].

3.2.5. Kinds of stimuli that can be attended

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 the subject of experiments and modeling.

In the Neurodynamical Model [169], locations are selected as points at the resolution of V1, but within each hypercolumn, features at a certain scale emerge as winners of the competition, thus in effect achieving both precise localization and varying sizes of the attentional focus. To mimic the observed faster processing of lower frequency stimuli, time constants are chosen such that the competition converges faster for them (note that this is very different from the "quick and dirty" processing proposed by [164] and [155] reviewed above).

In many cases we can see a marked difference between the theoretical capabilities of a model and those of its early implementations, differences that are sometimes addressed in later work. For example, while Selective Tuning as a theoretical model imposes no limits on the shape of the attentional focus, the implementation presented in [63] is based on rectangular patches of space at various scales and aspect ratios. More recently, [168] introduces a Θ-winner-take-all[3] algorithm that allows the shape of the focus to be driven by the input data, in what is an impressive demonstration of the ability of the model to perform its two main stated goals: identification and localization of motion stimuli in natural image sequences.

Another example, again illustrated using Selective Tuning [63] concerns the number of locations or objects that can be attended. Given that the nature of cortical connectivity, and especially the inhibitory lateral connections, is local, a global WTA competition is highly unlikely, so it is entirely possible that multiple winners can emerge out of the competition, and thus multiple attentional beams can operate in parallel, but

[3] See footnote on page 25.

this is avoided by implementing the decision in layers of extremely low resolution, effectively imposing global competition and restricting the system to a single attentional focus. It would be very interesting to see a demonstration of Selective Tuning where this limitation is removed, especially because some of the predictions of the model seem to require it. For example, some of the results presented by Cutzu and Tsotsos in support of Selective Tuning [142] seem to require a second attentional focus to be consistent with the theory.

A model that explicitly tries to address the lack of multiple attentional foci is [156], unfortunately the model appeals to a biologically implausible symbolic stage that represents a literal interpretation of the notion of "object files" and stores information such as position, size and trajectory for the selected objects.

The first model to implement object-based attention is MORSEL [161]. For more information see section 3.2.6. An interesting aspect of MORSEL is the pattern of connectivity within the "attentional mechanism" network (see above). Here units have local excitatory and distant inhibitory connections, which allows the network to converge into an "elastic" spotlight, with size and position dependent on the input image.

While not a computational model of visual attention per se, MAGIC, the system presented in [112] is an effort to simulate the results of Egly-type same-object advantage experiments. The hierarchical system learns feature grouping from labeled examples, and uses both firing rate and spike phase to encode the information. For a given neuron, the firing rate is a measure of the confidence in the grouping of the particular feature, while the phase correlation between features is used to indicate grouping. This information is correlated to psychophysical performance in Egly-type experiments that the authors present. The results seem to support the notion that strength of grouping can explain the observed same-object advantage.

The limitations on the resolution of attention presented by Intriligator and Cavanagh [176], and explained as limits imposed by the size of the receptive fields in LIP have not been addressed by the modeling community, but in the author's opinion a more plausible explanation could be proposed in the context of Selective Tuning [63] or Neurodynamical [169] Models. This may be caused by the size of the competitive circuits in V1 and/or the extent of the feedback connections. This is supported by a series of studies that find a very good similarity between the sizes of lateral connections in V1 and the size of the feedback projections of individual neurons in V2 with the limits found in attentional resolution [177].

3.2.6. Attention and recognition

Many modeling efforts separate attention and recognition, and even today, some researchers persist in this approach, e.g. Draper and Lionelle [153] declare that the purpose of a model of visual attention is to be "the front end to an appearance based object recognition system."

Probably the most blatant example of this approach is the Shifter Circuits Model [149, 150], 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 a cropped area of the input image has been selected, it is presented to an associative recognition network. The same approach is taken by the SCAN model [178] and by the Selective Attention for Identification Model (SAIM) [179].

MORSEL [161] integrates attentional selection into an object recognition network (in particular, stylized 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 miscombine letters to form words that are not present (similar to the "letter migration" phenomenon observed in perceptual studies [180]). 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. [181]. In this case, the saliency-based attentional system of Itti et al. [151, 182] operates in parallel to the hierarchical recognition system of Riesenhuber and Poggio [183], 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, high contrast paper-clip type objects on dark uniform backgrounds, 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 problems.

The diametrically opposite approach is the total integration of attention and object recognition, a solution pioneered by Fukushima's Neocognitron system [172]. 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 feedforward pathway, they follow the same route as the feedforward signals. If a feature is missing, the feedback is blocked, which causes a lowering of the detection threshold in the feedforward 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.

The object recognition capabilities of the Selective Tuning Model [63] are explored by Dolson [184]. In this work, simple object recognition is implemented as a process of reconstruction from parts, the top-level selection being considered a recognition hypothesis. The role of attention is to prune the processing hierarchy of all information that is not consistent with the hypothesis, thus validating (or invalidating) it.

Two different schemes of integrated attention and object recognition are investigated in the context of the Neurodynamical model. The first one, presented in [158] has been discussed above. In [169], attention is implemented through local competition biased by top-down connections, while object recognition is implemented in the feedforward 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.

While not object recognition in the traditional sense, Tsotsos et al. [168] presents an extension of the Selective Tuning model [63] 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, and 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.

One important and little understood aspect of the interaction between attention and object recognition is the fact that while certain very complicated stimuli can be recognized in the absence or near-absence of attention, simple stimuli like rotated T's and L's can not be discriminated without the full deployment of attention. Understanding this phenomenon can lead to significant insights into the mechanisms of object recognition and the binding problem, but none of the models reviewed address this issue beyond simply mentioning it.

The amount of research into the area of attentional object recognition demonstrates its importance and actuality, but with the exception of early results in multi-object recognition, and some purely technical contributions, our understanding of the interaction between attention and object recognition is very limited, and this is one of the most promising areas for computational modeling to make significant contributions.

4. Conclusions

This review discussed and classified the main theories and computational models of visual attention.

A number of areas in computer vision that take a similar approach have been briefly reviewed, followed by a brief presentation of the most relevant experimental paradigms and findings that have guided the development of computational neuroscience models of visual attention.

The key components of any complete theory of attention have been introduced, and, based on these, the major theories of primate visual attention have been presented in an original classification, qualitatively different, but not incompatible with the one proposed by Fernandez-Duque and Johnson [185]. With this foundation in place, the major classes of computational models of visual attention have been analyzed, and with emphasis on how the identified key components are addressed.

The important contributions that this paper makes are the identification of the many disparate components that fall under the definition of attention, the analysis of the main computational models of visual attention within this framework, and the original classification of visual attention theories.

The main benefit of this systematic analysis is its ability to identify areas that have received significant attention from the research community and, more importantly, areas where there are major open questions.

The modeling of saliency is the dominant theme of visual attention modeling research, probably due to a combination of historic, technical, and subjective reasons. Historically, the first major theory of visual attention, Feature Integration Theory [129] falls under this category, and several theories have developed these ideas with quite considerable success in describing psychophysics results (but offer little, if any, insight into neurophysiology). From a technical perspective, these theories are very easy to implement, and match well with the extensive computer science work on region of interest operators. Subjectively, the picture presented in the context of these theories and models is very attractive, being very simple and intuitive. The notion of a localized saliency map in the brain is very compatible with the modular and reductionist approach that has characterized much of neuroscience in the past.

Most theories and models limit top-down influences to a biasing role, modulating the combination of features into saliency maps or the competition between representations. While this may be the case, neurophysiology also shows that top-down influences play a crucial role in everything from figure-ground segregation to object recognition, and the fact that most of these effects are not present in anesthetized animals and in the absence of attention proves their very active role that modeling can not ignore.

Very little research has focused on attentional selection and filtering, and indeed, most (saliency based) models are too simplistic to even approach these subjects. This is the area in which dynamical models have the best chance to make a significant contribution, as they do not suffer from the limitations that characterize other models.

Inhibition of return and overt attention have long been subjects of computational modeling research, but rarely have biologically plausible mechanisms been integrated in visual attention models. This is beginning to change, and recent work has made significant contributions in this direction.

Similar to attentional selection and filtering, the issue of the kinds of stimuli that can be attended seems very much the domain of dynamical models. Recent attempts by saliency based models to approach this problem (in particular object recognition) have only served as reminders of their limitations, and while some of the results are technically sound, they greatly depart from biological plausibility, being in effect region of interest based systems.

Object recognition is the crown jewel of computer vision, and the fact that biologically plausible models seem unable to approach the level of performance of computer vision systems (let alone that of the primate brain) is a strong indication that much remains to be done in this area. Progress here is very likely to also contribute significantly to most of the other areas of research, especially in addressing questions about attentional filtering and the kinds of stimuli that can be attended.

In this respect, while this paper has identified open questions in all areas of research, object recognition stands out as probably the most important target 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 Michael Jenkin, Richard Wildes, Hugh Wilson, Eric Leung and Michael Tombu for valuable substantive and editorial comments. The authors alone are responsible for any omissions and errors.

Bibliography

1. Pashler, H.E., *The psychology of attention*. 1998, Cambridge, Mass.: MIT Press. xiv, 494 p.
2. Tsotsos, J.K., *Analyzing Vision at the Complexity Level*. Behavioral and Brain Sciences, 1990. **13**(3): p. 423-444.
3. Tsotsos, J.K. *Visual Attention: A Brief History of Computational Approaches*. in *Rosenon Workshop on computational Vision*. 2003. Stockholm, Sweden.
4. Bajcsy, R. *Active perception vs. passive perception*. in *3rd IEEE Workshop on Computer Vision: Representation and Control*. 1985.
5. Bajcsy, R.K., *Active Perception*. Proceedings of the IEEE, 1988. **76**(8).
6. Aloimonos, J.Y., Weiss, I., and Bandopadhay, A., *Active Vision*. International Journal on Computer Vision, 1987. **1**: p. 333-356.
7. Clark, J.J. and Ferrier, N.J. *Modal Control of Attentive Vision System*. in *2nd International Conference on Computer Vision*. 1988. Tampa, FL.
8. Pahlavan, K. and Eklundh, J.O. *Eye and Head-Eye System*. in *SPIE Applications of AI X: Machine Vision and Robotics*. 1992. Orlando, Fla.
9. Ferrier, N.J. and Clark, J.J., *The Harvard Binocular Head*. International Journal of Pattern Recognition and Artificial Intelligence, 1993: p. 9-31.
10. Brown, C.M., Coombs, D., and Soong, J., *Real-Time Smooth Pursuit Tracking*, in *Active vision*, A. Blake and A. Yuille, Editors. 1992, MIT Press. p. 123-136).
11. Clark, J.J. and Ferrier, N.J., *Attentive Visual Servo Control*, in *Active Vision*, A.B.a.A. Yuille, Editor. 1992, MIT Press. p. 137--154.
12. Coombs, D.J. and Brown, C.M., *Real-Time Binocular Smooth Pursuit*. International Journal of Computer Vision, 1993. **11**(2): p. 147-164.
13. Christensen, H.I., *A Low-Cost Robot Camera Head*. International Journal of Pattern Recognition and Artificial Intelligence, 1993. **7**(1): p. 69-87.
14. Milios, E., Jenkin, M., and Tsotsos, J.K., *Design and Performance of TRISH, a Binocular Robot Head with Torsional Eye Movements*. International Journal of Pattern Recognition and Artificial Intelligence, 1993. **7**(1): p. 51-68.
15. Swets, D.L. and Weng, J., *Hierarchical Discriminant Analysis for Image Retrieval*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1999. **21**(5): p. 386-401.
16. Behnke, S. and Karayiannis, N.B., *Competitive Neural Trees for Pattern Classification*. Ieee Transactions on Neural Networks, 1998. **9**(6): p. 1352-1369.
17. Becanovic, V., Kermit, M., and Eide, A.J., *Feature extraction from photographical images using a hybrid neural network*, in *Virtual Intelligence/Dynamic Neural Networks*, L.e. al., Editor. 1998: Stockholm, Sweden. p. 351-361.
18. Rao, R.P.N. and Ballard, D.H., *An Active Vision Architecture based on Iconic Representations*. Artificial Intelligence, 1995. **78**: p. 461-505.

19. Krotkov, E., Henriksen, K., and Kories, R., *Stereo Ranging from Verging Cameras*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1990. **12**: p. 1200-1205.
20. Ballard, D.H. and Ozcandarli, A. *Eye Fixation and Early Vision: Kinematic Depth*. in *IEEE 2nd Intl. Conf. on Comp. Vision*. 1988. Tarpon Springs, Fla.
21. Brown, C., *Prediction and Cooperation in Gaze Control*. Biological Cybernetics, 1990. **63**.
22. Das, S., Abbott, A.L., and Ahuja, N. *Surface Reconstruction from Focus and Stereo*. in *5th International Conference on Image Analysis and Processing*. 1989. Positano, Italy.
23. Ahuja, N. and Abbott, A.L., *Active Stereo: Integrating Disparity, Vergence, Focus, Aperture, and Calibration for Surface Estimation*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1993. **15**(10): p. 1007-1029.
24. Das, S. and Ahuja, N. *Multiresolution Image Acquisition and Surface Reconstruction*. in *3rd International Conference on Computer Vision*. 1990. Osaka, Japan.
25. Klarquist, W.N. and Bovik, A.C., *FOVEA: A foveated vergent active stereo system for dynamic three-dimensional scene recovery*. IEEE Transactions on Robotics and Automation, 1998. **14**(5): p. 755-770.
26. Sarkar, S. and Boyer, K.L., *Perceptual Organization in Computer Vision - a Review and a Proposal for a Classifactory Structure*. Ieee Transactions on Systems Man and Cybernetics, 1993. **23**(2): p. 382-399.
27. Engbers, E.A. and Smeulders, A.W.M., *Design considerations for generic grouping in vision*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2003. **25**(4): p. 445-457.
28. Lowe, D.G., *Perceptual organization and Visual Recognition*. 1985, Boston, MA: Kluwer.
29. Tenenbaum, J.M. and Witkin, A., *Perceptual Organization as Building-Blocks for Vision*. Journal of the Optical Society of America a-Optics Image Science and Vision, 1984. **1**(12): p. 1216-1216.
30. Stevens, K.A., *Computation of Locally Parallel Structure*. Biological Cybernetics, 1978. **29**(1): p. 19-28.
31. Zucker, S.W., *Computational and Psychophysical Experiments in Grouping: Early Orientation Selection*, in *Human and Machine Vision*, J. Beck, B. Hope, and A. Rosenfeld, Editors. 1983, Academic Press. p. 545-567.
32. Tang, B., Sapiro, G., and Caselles, V., *Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case*. International Journal of Computer Vision, 2000. **36**(2): p. 149-161.
33. Sochen, N.A. and Kimmel, R., *Combing a porcupine via stereographic direction diffusion*. Scale-Space and Morphology in Computer Vision, Proceedings, 2001. **2106**: p. 308-316.
34. Kimmel, R. and Sochen, N., *Orientation diffusion or how to comb a porcupine*. Journal of Visual Communication and Image Representation, 2002. **13**(1-2): p. 238-248.

35. Ben-Shahar, O. and Zucker, S.W., *The perceptual organization of texture flow: A contextual inference approach*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2003. **25**(4): p. 401-417.
36. Little, J., Bulthoff, H., and Poggio, T. *Parallel Optical Flow Using Local Voting*. in *International Conference on Computer Vision*. 1988.
37. Medioni, G., Lee, M.-S., and Tang, C.-K., *A Computational Framework for Segmentation and Grouping*. 2000: Elsevier.
38. Gaucher, L. and Medioni, G. *Accurate Motion Flow Estimation with Discontinuities*. in *International Conference on Computer Vision*. 1999.
39. Wu, Y., Kanade, T., Cohn, J., and Li, C. *Optic Flow Estimation Using Wavelet Motion Model*. in *International Conference on Computer Vision*. 1998.
40. Shi, J. and Malik, J. *Motion Segmentation and Tracking Using Normalized Cuts*. in *International Conference on Computer Vision*. 1998.
41. Fleet, D., Black, M., and Jepson, A. *Motion Feature Detection Using Steerable Flow Fields*. 1998.
42. Haralick, R.M. and Shapiro, L.G., *Computer and robot vision*. 1992, Reading, Mass.: Addison-Wesley Pub. Co.
43. Moravec, H.P. *Towards automatic visual obstacle avoidance*. in *International Joint Conference on Artificial Intelligence*. 1977.
44. Moravec, H.P. *Rover visual obstacle avoidance*. in *International Joint Conference on Artificial Intelligence*. 1981.
45. Harris, C., *Geometry from visual motion*, in *Active Vision*, A. Blake and A. Yuille, Editors. 1992, MIT Press. p. 263-284.
46. Harris, C. and Stephens, M. *A combined corner and edge detector*. in *Fourth Alvey Vision Conference*. 1988. Manchester, U.K.
47. Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., and Kubler, O., *Simulation of neural contour mechanisms: from simple to end-stopped cells*. Vision Research, 1992. **32**(5): p. 963-81.
48. Forstner, W. *A framework for low level feature extraction*. in *European Conference on Computer Vision 2*. 1994.
49. Zhang, Z.Y., Deriche, R., Faugeras, O., and Luong, Q.T., *A Robust Technique for Matching 2 Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry*. Artificial intelligence, 1995. **78**(1-2): p. 87-119.
50. Torr, P., *Motion Segmentation and Outlier Detection*, in *Dept. of Engineering Science*. 1995, Oxford University: Oxford, U.K.
51. Schmid, C. and Mohr, R., *Local grayvalue invariants for image retrieval*. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1997. **19**(5): p. 530-535.
52. Lindeberg, T., *Feature detection with automatic scale selection*. International Journal of Computer Vision, 1998. **30**(2): p. 79-116.
53. Mikolajczyk, K. and Schmid, C. *Indexing based on scale invariant interest points*. in *ICCV*. 2001.
54. Lowe, D.G. *Object recognition from local scale-invariant features*. in *International Conference on Computer Vision*. 1999. Corfu, Greece.
55. Schmid, C., Mohr, R., and Bauckhage, C., *Evaluation of interest point detectors*. International Journal of Computer Vision, 2000. **37**(2): p. 151-172.

56. Tian, Q., Sebe, N., Lew, M.S., Loupias, E., and Huang, T.S., *Image retrieval using wavelet-based salient points*. Journal of Electronic Imaging, 2001. **10**(4): p. 835-849.

57. Sebe, N., Tian, Q., Loupias, E., Lew, M.S., and Huang, T.S., *Evaluation of salient point techniques*. Image and Vision Computing, 2003. **21**(13-14): p. 1087-1095.

58. Loupias, E., Sebe, N., Bres, S., and Jolion, J.-M. *Wavelet-based salient points for image retrieval*. in *International Conference on Image Processing*. 2000.

59. Loupias, E. and Bres, S., *Key points-based indexing for pre-attentive similarities: The KIWI system*. Pattern Analysis and Applications, 2001. **4**(2-3): p. 200-214.

60. Burt, P. *Attention mechanisms for vision in a dynamic world*. in *Ninth International Conference on Pattern Recognition*. 1988. Beijing, China.

61. Burt, P., *A Pyramid Framework for Real-Time Computer Vision*, in *Foundations of Image Understanding*, L.S. Davis, Editor. 2001, Kluwer International. p. 349-380.

62. Conception, V. and Wechsler, H., *Detection and localization of objects in time-varying imagery using attention, representation and memory pyramids*. Pattern Recognition, 1996. **29**(9): p. 1543-1557.

63. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y.H., Davis, N., and Nuflo, F., *Modeling Visual-Attention Via Selective Tuning*. Artificial intelligence, 1995. **78**(1-2): p. 507-545.

64. Howarth, R.J. and Buxton, H. *Selective Attention in Dynamic Vision*. in *IJCAI93*. 1993.

65. Baluja, S. and Pomerleau, D., *Dynamic relevance: vision-based focus of attention using artificial neural networks*. Artificial intelligence, 1997. **97**(1-2): p. 381-395.

66. Sela, G. and Levine, M.D., *Real-time attention for robotic vision*. Real-Time Imaging, 1997. **3**(3): p. 173-194.

67. Gallet, O., Gaussier, P., and Cocquerez, J.-P. *A model of the visual attention to speed up image analysis*. in *International Conference on Image Processing*. 1998.

68. Gaussier, P., Joulain, C., Banquet, J.P., Lepretre, S., and Revel, A., *The visual homing problem: An example of robotics/biology cross fertilization*. Robotics and Autonomous Systems, 2000. **30**(1-2): p. 155-180.

69. Toyama, K. and Hager, G.D., *Incremental focus of attention for robust vision-based tracking*. International Journal of Computer Vision, 1999. **35**(1): p. 45-63.

70. Sun, Y.R. and Fisher, R., *Object-based visual attention for computer vision*. Artificial intelligence, 2003. **146**(1): p. 77-123.

71. Zabrodsky, H. and Peleg, S., *Attentive transmission*. Journal of Visual Communications and Image Representation, 1990. **1**: p. 189-198.

72. Wolfe, J.M., *Visual Search*, in *Attention*, H. Pashler, Editor. 1998, Taylor & Francis: Philadelphia.

73. Quinlan, P.T. and Humphreys, G.W., *Visual search for targets defined by combinations of color, shape, and size: an examination of the task constraints on feature and conjunction searches*. Perception & Psychophysics, 1987. **41**(5): p. 455-72.

74. Rosenholtz, R., *Search asymmetries? What search asymmetries?* Perception & Psychophysics, 2001. **63**(3): p. 476-89.

75. Wolfe, J.M., *Moving towards solutions to some enduring controversies in visual search.* Trends Cogn Sci, 2003. **7**(2): p. 70-76.
76. Treisman, A.M. and Gelade, G., *Feature-Integration Theory of Attention.* Cognitive Psychology, 1980. **12**(1): p. 97-136.
77. Koch, C. and Ullman, S., *Shifts in selective visual attention: towards the underlying neural circuitry.* Human Neurobiology, 1985. **4**(4): p. 219-27.
78. Kusunoki, M., Gottlieb, J., and Goldberg, M.E., *The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance.* Vision Research, 2000. **40**(10-12): p. 1459-68.
79. Bisley, J.W. and Goldberg, M.E., *Neuronal activity in the lateral intraparietal area and spatial attention.* Science, 2003. **299**(5603): p. 81-86.
80. Kapadia, M.K., Ito, M., Gilbert, C.D., and Westheimer, G., *Improvement in Visual Sensitivity by Changes in Local Context - Parallel Studies in Human Observers and in V1 of Alert Monkeys.* Neuron, 1995. **15**(4): p. 843-856.
81. Gallant, J.L., Van Essen, D.C., and Nothdurft, H.C., *Two-dimensional and three-dimensional texture processing in visual cortex of the macaque monkey*, in *Early vision and beyond*, T.V. Papathomas, Editor. 1995, The MIT Press: Cambridge, MA, US. p. 89-98.
82. Nothdurft, H.C., Gallant, J.L., and Van Essen, D.C., *Response profiles to texture border patterns in area V1.* Visual Neuroscience, 2000. **17**(3): p. 421-436.
83. Grossberg, S. and Mingolla, E., *Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations.* Perception & Psychophysics, 1985. **38**(2): p. 141-71.
84. Li, Z., *Pre-attentive segmentation in the primary visual cortex.* Spat Vis, 2000. **13**(1): p. 25-50.
85. Wilson, H.R., *Non-Fourier Cortical Processes in Texture, Form, and Motion Perception*, in *Cerebral Cortex*, P.S. Ulinski and E.G. Jones, Editors. 1999, Kluwer Academic/ Plenum Publishers: New York.
86. Lin, L.M. and Wilson, H.R., *Fourier and non-Fourier pattern discrimination compared.* Vision Research, 1996. **36**(13): p. 1907-18.
87. Reynolds, J.H., Pasternak, T., and Desimone, R., *Attention increases sensitivity of V4 neurons.* Neuron, 2000. **26**(3): p. 703-14.
88. Martinez-Trujillo, J.C. and Treue, S., *Feature-based attention increases the selectivity of population responses in primate visual cortex.* Curr Biol, 2004. **14**(9): p. 744-51.
89. Rothenstein, D.A., Tsotsos, J.K., and Martinez-Trujillo, J.C. *Attending to motion, how much can we see when removing attention from targets?* in *CVR Conference.* 2003. Toronto, Ontario.
90. McAdams, C.J. and Maunsell, J.H., *Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4.* Journal of Neuroscience, 1999. **19**(1): p. 431-41.
91. Treue, S. and Martinez Trujillo, J.C., *Feature-based attention influences motion processing gain in macaque visual cortex.* Nature, 1999. **399**(6736): p. 575-9.
92. Moran, J. and Desimone, R., *Selective attention gates visual processing in the extrastriate cortex.* Science, 1985. **229**(4715): p. 782-4.

93. Reynolds, J.H., Chelazzi, L., and Desimone, R., *Competitive mechanisms subserve attention in macaque areas V2 and V4*. Journal of Neuroscience, 1999. **19**(5): p. 1736-53.
94. Hoffman, J.E. and Subramaniam, B., *The role of visual attention in saccadic eye movements*. Perception & Psychophysics, 1995. **57**(6): p. 787-95.
95. Sheliga, B.M., Riggio, L., and Rizzolatti, G., *Orienting of attention and eye movements*. Exp Brain Res, 1994. **98**(3): p. 507-22.
96. Kowler, E., Anderson, E., Dosher, B., and Blaser, E., *The role of attention in the programming of saccades*. Vision Research, 1995. **35**(13): p. 1897-916.
97. Colby, C.L. and Goldberg, M.E., *Space and attention in parietal cortex*. Annual Review of Neuroscience, 1999. **22**: p. 319-349.
98. Andersen, R.A., Bracewell, R.M., Barash, S., Gnadt, J.W., and Fogassi, L., *Eye Position Effects on Visual, Memory, and Saccade-Related Activity in Areas Lip and 7a of Macaque*. Journal of Neuroscience, 1990. **10**(4): p. 1176-1196.
99. Maljkovic, V. and Nakayama, K., *Priming of pop-out: I. Role of features*. Mem Cognit, 1994. **22**(6): p. 657-72.
100. Chun, M.M. and Jiang, Y., *Contextual cueing: Implicit learning and memory of visual context guides spatial attention*. Cognitive Psychology, 1998. **36**(1): p. 28-71.
101. Blaser, E., Sperling, G., and Lu, Z.L., *Measuring the amplification of attention*. Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(20): p. 11681-11686.
102. Chelazzi, L., Duncan, J., Miller, E.K., and Desimone, R., *Responses of neurons in inferior temporal cortex during memory-guided visual search*. Journal of Neurophysiology, 1998. **80**(6): p. 2918-2940.
103. Posner, M.I., Snyder, C.R., and Davidson, B.J., *Attention and the detection of signals*. J Exp Psychol, 1980. **109**(2): p. 160-74.
104. Eriksen, C.W. and St James, J.D., *Visual attention within and around the field of focal attention: a zoom lens model*. Perception & Psychophysics, 1986. **40**(4): p. 225-40.
105. Downing, C. and Pinker, S., *The spatial structure of visual attention*, in *Attention and performance*, M. Posner and O.S.M. Marin, Editors. 1985, Erlbaum: London. p. 171-187.
106. Neisser, U., *Cognitive psychology*. 1967, New York,: Appleton-Century-Crofts. xi, 351 p.
107. Neisser, U. and Becklen, R., *Selective looking: Attending to visually specified events*. Cognitive Psychology, 1975. **7**(4): p. 480-494.
108. Neisser, U., *The control of information pickup in selective looking*, in *Perception and its development*, A. Pick, Editor. 1979, Erlbaum: Hillsdale, NJ. p. 201-219.
109. Most, S.B., Simons, D.J., Scholl, B.J., Jimenez, R., Clifford, E., and Chabris, C.F., *How not to be seen: the contribution of similarity and selective ignoring to sustained inattentional blindness*. Psychol Sci, 2001. **12**(1): p. 9-17.
110. Simons, D.J. and Chabris, C.F., *Gorillas in our midst: sustained inattentional blindness for dynamic events*. Perception, 1999. **28**(9): p. 1059-74.

111. Egly, R., Driver, J., and Rafal, R.D., *Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects*. J Exp Psychol Gen, 1994. **123**(2): p. 161-77.

112. Behrmann, M., Zemel, R.S., and Mozer, M.C., *Object-based attention and occlusion: Evidence from normal participants and a computational model.* Journal of Experimental Psychology: Human Perception and Performance, 1998. **24**(4): p. 1011-1036.

113. Moore, C.M., Yantis, S., and Vaughan, B., *Object-based visual selection: Evidence from perceptual completion.* Psychological Science, 1998. **9**(2): p. 104-110.

114. Robertson, I. and Marshall, J., *Unilateral neglect: clinical and experimental studies*. 1993: Erlbaum.

115. Rafal, R.D., *Neglect*, in *The attentive brain*, R. Parasuraman, Editor. 1998, MIT Press: Cambridge, MA. p. 489-525.

116. Rafal, R.D., *Balint syndrome*, in *Behavioral neurology and neuropsychology*, T. Feinberg and M. Farah, Editors. 1997, McGraw-Hill: New York. p. 337-356.

117. Caramazza, A. and Hillis, A.E., *Levels of representation, coordinate frames, and unilateral neglect*. Cognitive Neuropsychology, 1990. **7**(5/6): p. 391-445.

118. Behrmann, M. and Tipper, S.P., *Object-Based Attentional Mechanisms - Evidence from Patients with Unilateral Neglect*. Attention and Performance Xv, 1994. **15**: p. 351-375.

119. Behrmann, M. and Tipper, S.P., *Attention accesses multiple reference frames: evidence from visual neglect*. J Exp Psychol Hum Percept Perform, 1999. **25**(1): p. 83-101.

120. Scholl, B.J., *Objects and attention: the state of the art*. Cognition, 2001. **80**(1-2): p. 1-46.

121. Pashler, H., ed. *Attention.* 1998, Taylor & Francis: Philadelphia.

122. Newell, A., *Unified theories of cognition.* 1990, Cambridge, MA: Harvard University Press. 530.

123. Klein, R.M., *Inhibition of return.* Trends Cogn Sci, 2000. **4**(4): p. 138-147.

124. Horowitz, T.S. and Wolfe, J.M., *Visual search has no memory*. Nature, 1998. **394**(6693): p. 575-7.

125. Broadbent, D.E., *Perception and communication*. 1958, New York,: Pergamon Press. 338 p.

126. Treisman, A., *Contextual cues in selective listening*. Quarterly Journal of Experimental Psychology, 1960. **12**: p. 242-248.

127. Deutsch, J.A. and Deutsch, D., *Attention - Some Theoretical Considerations*. Psychological Review, 1963. **70**(1): p. 80-90.

128. Norman, D.A., *Toward a theory of memory and attention*. Psychological Review, 1968. **75**(6): p. 522 536.

129. Treisman, A., *Selective attention in man*. British Medical Bulletin, 1964. **20**: p. 12-16.

130. Johnston, W.A. and Heinz, S.P., *Flexibility and Capacity Demands of Attention.* Journal of Experimental Psychology-General, 1978. **107**(4): p. 420-435.

131. Johnston, W.A. and Heinz, S.P., *Depth of nontarget processing in an attention task*. J Exp Psychol Hum Percept Perform, 1979. **5**(1): p. 168-75.

132. Wolfe, J.M., Cave, K.R., and Franzel, S.L., *Guided search: an alternative to the feature integration model for visual search*. J Exp Psychol Hum Percept Perform, 1989. **15**(3): p. 419-33.

133. Wolfe, J.M., *Guided search 2.0: A revised model of visual search*. Psychonomic Bulletin and Review, 1994. **1**(2): p. 202-238.

134. Wolfe, J.M., *Extending Guided Search: Why Guided Search needs a preattentive "item map."* in *Converging operations in the study of visual selective attention*, A.F. Kramer, Editor. 1996, American Psychological Association: Washington, DC, US. p. 247-270.

135. Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L.G., *Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI*. Science, 1998. **282**(5386): p. 108-11.

136. Britten, K.H., *Cortical neurophysiology - Attention is everywhere*. Nature, 1996. **382**(6591): p. 497-498.

137. Vanduffel, W., Tootell, R.B.H., and Orban, G.A., *Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system*. Cerebral Cortex, 2000. **10**(2): p. 109-126.

138. Mehta, A.D., Ulbert, I., and Schroeder, C.E., *Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas*. Cerebral Cortex, 2000. **10**(4): p. 343-58.

139. Desimone, R. and Duncan, J., *Neural Mechanisms of Selective Visual-Attention*. Annual Review of Neuroscience, 1995. **18**: p. 193-222.

140. Caputo, G. and Guerra, S., *Attentional selection by distracter suppression*. Vision Research, 1998. **38**(5): p. 669-689.

141. Bahcall, D.O. and Kowler, E., *Attentional interference at small spatial separations*. Vision Research, 1999. **39**(1): p. 71-86.

142. Cutzu, F. and Tsotsos, J.K., *The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item*. Vision Research, 2003. **43**(2): p. 205-219.

143. Joseph, J.S., Chun, M.M., and Nakayama, K., *Attentional requirements in a 'preattentive' feature search task*. Nature, 1997. **387**(6635): p. 805-7.

144. Rothenstein, A.L., Martinez-Trujillo, J.C., Treue, S., Tsotsos, J.K., and Wilson, H.R. *Modeling Attentional Effects in Cortical Areas MT and MST of the Macaque Monkey Through Feedback Loops*. in *Society for Neuroscience Annual Meeting*. 2002. Orlando, Florida.

145. Bundensen, C., *Visual Selective Attention: Outlines of a Choice Model, a Race Model and a Computational Theory*. Visual Cognition, 1998. **5**(1/2): p. 287-309.

146. Eckstein, M.P., Thomas, J.P., Palmer, J., and Shimozaki, S.S., *A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays*. Perception & Psychophysics, 2000. **62**(3): p. 425-451.

147. Verghese, P., *Visual search and attention: a signal detection theory approach*. Neuron, 2001. **31**(4): p. 523-35.

148. Eckstein, M.P., Shimozaki, S.S., and Abbey, C.K., *The footprints of visual attention in the Posner cueing paradigm revealed by classification images*. J Vis, 2002. **2**(1): p. 25-45.

149. Olshausen, B.A., Anderson, C.H., and Van Essen, D.C., *A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information*. Journal of Neuroscience, 1993. **13**(11): p. 4700-19.

150. Anderson, C. and Van Essen, D., *Shifter Circuits: a computational strategy for dynamic aspects of visual processing*. Proc. Natl. Academy Sci. USA, 1987. **84**: p. 6297-6301.

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

152. Itti, L. and Koch, C. *A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems*. in *SPIE Human Vision and Electronic Imaging-IV*. 1999. San Jose, CA.

153. Draper, B. and Lionelle, A. *Evaluation of Selective Attention under Similarity Transforms*. in *International workshop on attention and performance in computer vision (WACPV 2003)*. 2003. Graz, Austria.

154. vandeLaar, P., Heskes, T., and Gielen, S., *Task-dependent learning of attention*. Neural Networks, 1997. **10**(6): p. 981-992.

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

156. Backer, G. and Mertsching, B. *Two Selection Stages Provide Efficient Object-based Attentional Control for Dynamic Vision*. in *International workshop on attention and performance in computer vision (WACPV 2003)*. 2003. Graz, Austria.

157. Gottlieb, J.P., Kusunoki, M., and Goldberg, M.E., *The representation of visual salience in monkey parietal cortex*. Nature, 1998. **391**(6666): p. 481-4.

158. Deco, G. and Zihl, J., *A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system*. J Comput Neurosci, 2001. **10**(3): p. 231-53.

159. Grossberg, S., *How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex*. Spatial Vision, 1999. **12**(2): p. 163-185.

160. Li, Z., *Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex*. Network, 1999. **10**(2): p. 187-212.

161. Mozer, M.C., *The perception of multiple objects : a connectionist approach*. Neural network modeling and connectionism. 1991, Cambridge, Mass.: MIT Press. 217 p.

162. Zaharescu, A., Rothenstein, A.L., and Tsotsos, J.K. *Towards a Biologically Plausible Active Visual Search Model*. in *WACPV*. 2004. Prague, Czech Republic.

163. Motter, B.C., *Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli*. Journal of Neurophysiology, 1993. **70**(3): p. 909-19.

164. Oliva, A., Torralba, A., Castelhano, M.S., and Henderson, J.M. *Top-Down control of visual attention in object detection*. in *IEEE International Conference on Image Processing*. 2003. Barcelona, Spain.

165. Torralba, A. and Oliva, A., *Statistics of natural image categories*. Network-Computation in Neural Systems, 2003. **14**(3): p. 391-412.

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

167. Wixson, L., *Gaze selection for visual search*. 1994, University of Rochester Dept. of Computer Science: Rochester, N.Y. p. xiv, 155 p.

168. Tsotsos, J.K., Pomplun, M., Liu, Y., Martinez-Trujillo, J.C., and Simine, E. *Attending to Motion: Localizing and Labeling Simple Motion Patterns in Image Sequences*. in *Conference on Biologically-Motivated Computer Vision*. 2002. Tuebingen, Germany.

169. Rolls, E.T. and Deco, G., *Computational neuroscience of vision*. 2002, Oxford ; New York: Oxford University Press. xviii, 569 p.

170. Slotnick, S.D., Schwarzbach, J., and Yantis, S., *Attentional inhibition of visual processing in human striate and extrastriate cortex*. Neuroimage, 2003. **19**(4): p. 1602-11.

171. Kristjansson, A. and Nakayama, K., *The attentional blink in space and time*. Vision Research, 2002. **42**(17): p. 2039-50.

172. Fukushima, K., Imagawa, T., and Ashida, E. *Character recognition with selective attention*. in *International Joint Conference on Neural Networks*. 1991. Seattle.

173. Wilson, H.R., *Spikes, decisions, and actions : the dynamical foundations of neuroscience*. 1999, Oxford ; New York: Oxford University Press.

174. Pouget, A. and Sejnowski, T.J., *Dynamical Remapping*, in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib, Editor. 1995, MIT Press: Boston. p. 335-338.

175. Lanyon, L.J. and Denham, S.L., *A biased competition computational model of spatial and object-based attention mediating active visual search*. Neurocomputing, 2004. **58-60**: p. 655-662.

176. Intriligator, J. and Cavanagh, P., *The spatial resolution of visual attention*. Cognit Psychol, 2001. **43**(3): p. 171-216.

177. Angelucci, A., Levitt, J.B., Walton, E.J.S., Hupe, J.M., Bullier, J., and Lund, J.S., *Circuits for local and global signal integration in primary visual cortex*. Journal of Neuroscience, 2002. **22**(19): p. 8633-8646.

178. Postma, E.O., vandenHerik, H.J., and Hudson, P.T.W., *SCAN: A scalable model of attentional selection*. Neural Networks, 1997. **10**(6): p. 993-1015.

179. Heinke, D. and Humphreys, G.W. *SAIM: A Model of Visual Attention and Neglect*. in *7th International Conference on Artificial Neural Networks*. 1997. Lausanne, Switzerland: Springer Verlag.

180. Mozer, M.C., *Letter Migration in Word Perception*. Journal of Experimental Psychology-Human Perception and Performance, 1983. **9**(4): p. 531-546.

181. Walther, D., Itti, L., Riesenhuber, M., Poggio, T., and Koch, C., *Attentional selection for object recognition - A gentle way*. Biologically Motivated Computer Vision, Proceedings, 2002. **2525**: p. 472-479.

182. Itti, L. and Koch, C., *A saliency-based search mechanism for overt and covert shifts of visual attention*. Vision Research, 2000. **40**(10-12): p. 1489-506.

183. Riesenhuber, M. and Poggio, T., *Hierarchical models of object recognition in cortex*. Nature Neuroscience, 1999. **2**(11): p. 1019-1025.

184. Dolson, D.C., *Attentive object recognition in the selective tuning network*, in *Department of Computer Science*. 1997, University of Toronto: Toronto. p. 163.

185. Fernandez-Duque, D. and Johnson, M.L., *Cause and effect theories of attention: The role of conceptual metaphors*. Review of General Psychology, 2002. **6**(2): p. 153-165.