Computational Cognitive Neuroscience and Its Applications

Laurent Itti
University of Southern California

Introduction and motivation

A number of tasks which can be effortlessly achieved by humans and other animals have until recently remained seemingly intractable for computers. Striking examples of such tasks exist in the visual and auditory domains. These include recognizing the face of a friend in a photograph, understanding whether a new, never-seen object is a car or an airplane, quickly finding objects or persons of interest like one's child in a crowd, understanding fluent speech while also deciphering the emotional state of the speaker, or reading cursive handwriting. In fact, several of these tasks have become the hallmark of human intelligence, while other, seemingly more complex and more cognitively involved tasks, such as playing checkers or chess, solving differential equations, or proving theorems have been mastered by machines to a reasonable degree (Samuel, 1959; Newell & Simon, 1972). An everyday demonstration of this state of affairs is the use of simple image, character, or sound recognition in CAPTCHA tests (Completely Automated Public Turing Tests to Tell Computers and Humans Apart) used by many web sites to ensure that a human, rather than a software robot, is accessing the site (CAPTCHA tests, for example, are used by web sites providing free email accounts to registered users, and are a simple yet imperfect way to prevent spammers from opening thousands of email accounts using an automated script).

To some extent, these machine-intractable tasks are the cause for our falling short on the early promises made in the 1950s by the founders of Artificial Intelligence, Computer Vision, and Robotics (Minsky, 1961). Although tremendous progress has been made in just a half century, and one is beginning to see cars that can drive on their own or robots that vacuum the floor without human supervision, such machines have not yet reached mainstream adoption and remain highly limited in their ability to interact with the real world. Although in the early years one could blame the poor performance of machines on limitations in computing resources, rapid advances in microelectronics have now rendered such excuses less believable. The core of the problem is not only how much computing cycles one may have to perform a task, but how those cycles are used, in what kind of algorithm and of computing paradigm.

For biological systems, interacting with the visual world, in particular through vision, audition, and other senses, is key to survival. Essential tasks like locating and identifying potential prey, predators, or mates must be performed quickly and reliably if an animal is to stay alive. Taking inspiration from nature, recent work in computational neuroscience has hence started to devise a new breed of algorithms, which can be more flexible, robust, and adaptive when confronted with the complexities of the real world. I here focus on describing recent progress with a few simple examples of such algorithms, concerned with directing attention towards interesting locations in a visual scene, so as to concentrate the deployment of computing resources primarily onto these locations.

Modeling visual attention

Positively identifying any and all interesting targets in one's visual field has prohibitive computational complexity, making it a daunting task even for the most sophisticated

biological brains (Tsotsos, 1991). One solution, adopted by primates and many other animals, is to break down the visual analysis of the entire field of view into smaller regions, each of which is easier to analyze and can be processed in turn. This serialization of visual scene analysis is operationalized through mechanisms of visual attention: A common (athough somewhat inaccurate) metaphor for attention is that of a virtual "spotlight," shifting towards and highlighting different sub-regions of the visual world, so that one region at a time can be subjected to more detailed visual analysis (Treisman & Gelade, 1980; Crick, 1984; Weichselgartner & Sperling, 1987). The central problem in attention research then becomes how to best direct this spotlight towards the most interesting and behaviorally relevant visual locations. Simple strategies, like constantly scanning the visual field from left to right and from top to bottom, like many computer algorithms do, may be too slow for situations where survival depends on reacting quickly. Recent progress in computational neuroscience has proposed a number of new biologically-inspired algorithms which implement more efficient strategies. These algorithms usually distinguish between a so-called "bottom-up" drive of attention towards conspicuous or "salient" locations in the visual field, and a volitional and task-dependent so-called "top-down" drive of attention toward behaviorally relevant scene elements (Desimone & Duncan, 1995; Itti & Koch, 2001). Simple examples of bottom-up salient and top-down relevant items are shown in Figure 1.

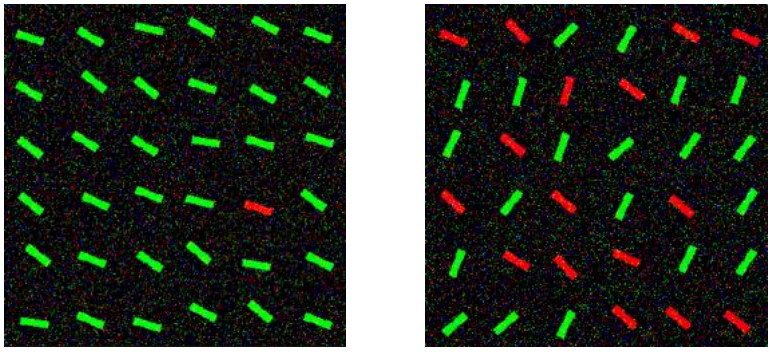

Figure 1: (left) Example where one item (a red and roughly horizontal bar) in an array of items is highly salient and immediately and effortlessly grabs visual attention attention in a bottom-up, image-driven manner. (right) Example where a similar item (a red and roughly vertical bar) is not salient but may be behaviorally relevant if your task is to find it as quickly as possible; top-down, volition-driven mechanisms, must be deployed to initiate a search for the item. (Also see Treisman & Gelade, 1980).

Koch and Ullman (1985) introduced the idea of a saliency map to accomplish preattentive selection in the primate brain. This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

Later research has further elucidated the basic principles behind computing salience (Figure 2). One important principle is the detection of locations whose local visual statistics significantly differ from the surrounding image statistics, along some dimension or combination of dimensions which are currently relevant to the subjective observer (Itti et al., 1998; Itti & Koch, 2001). This significant difference could be in a number of

simple visual feature dimensions which are believed to be represented in the early stages of cortical visual processing: color, edge orientation, luminance, or motion direction (Treisman & Gelade, 1980; Itti & Koch, 2001). Wolfe and Horowith (2004) provide a very nice review of which elementary visual features may strongly contribute to visual salience and guide visual search.

Two mathematical constructs can be derived from electrophysiological recordings in living brains, which shed light onto how this detection of statistical odd-man-out may be carried out. First, early visual neurons often have center-surround receptive field structures, by which the neuron's view of the world consists of two antagonistic sub-regions of the visual field, a small central region which drives the neuron in one direction (e.g., excites the neuron when a bright pattern is presented) and a larger, concentric surround region which drives the neuron in the opposite direction (e.g., inhibits the neuron when a bright pattern is presented; Kuffler, 1953). In later processing stages, this simple concentric center-surround receptive field structure is replaced by more complex types of differential operators, such as Gabor receptive fields sensitive to the presence of oriented line segments (Hubel & Wiesel, 1962). In addition, more recent research has unraveled how neurons with similar stimulus preferences but sensitive to different locations in the visual field may interact. In particular, neurons with sensitivities to similar patterns tend to inhibit each other, a phenomenon known as non-classical surround inhibition (Allman et al., 1985; Cannon & Fullenkamp, 1991; Sillito et al., 1995). Taken together, these basic principles suggest ways by which detecting an odd-man-out, or significantly different item in a display, can be achieved in a very economical (in terms of neural hardware) yet robust manner.

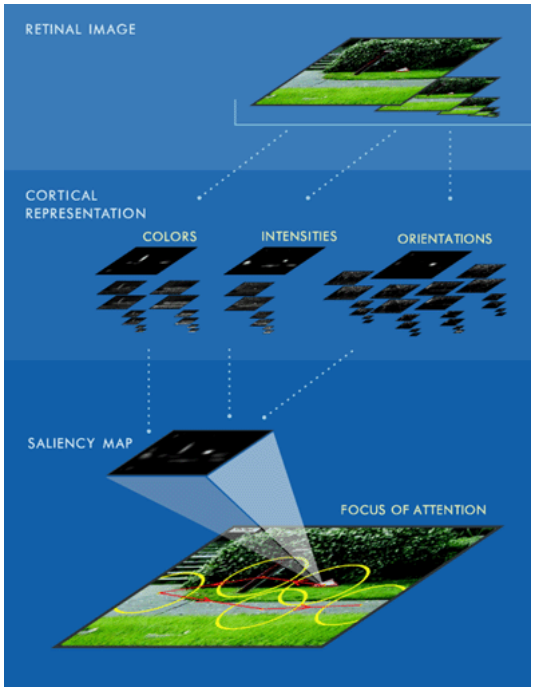

Figure 2: Architecture for computing saliency and directing attention towards conspicuous locations in a scene. Incoming visual input (top) is processes at several spatial scales along a number of basic feature dimensions, including color, luminance intensity, and local edge orientation. Center-surround operations as well as non-classical surround inhibition within each resulting "feature map" enhance the neural representation of locations which significantly stand out from their neighbors. All feature maps feed into a single saliency map which topographically represents salience irrespectively of features. Finally, attention is first directed to the most salient location in the image, and subsequently to less salient locations.

Many computational models of human visual search have embraced the idea of a saliency map under different guises (Treisman, 1988; Wolfe, 1994; Niebur & Koch, 1996; Itti &

Koch, 2000). The appeal of an explicit saliency map is the relatively straightforward manner in which it allows the input from multiple, quasi-independent feature maps to be combined and to give rise to a single output: The next location to be attended. Related formulations of this basic principle have been expressed in slightly different terms, including defining salient locations as those which contain spatial outliers (Rosenholtz, 1999), which may be more informative in Shannon's sense (Bruce & Tsotsos, 2006), or, in a more general formulation, which may be more surprising in a Bayesian sense (Itti & Baldi, 2006). Electrophysiological evidence points to the existence of several neuronal maps, in the pulvinar, the superior colliculus and the intraparietal sulcus, which appear to specifically encode for the saliency of a visual stimulus (Robinson & Petersen, 1992; Gottlieb et al., 1998; Colby & Goldberg, 1999).

Fully implemented computational models of attention, although relatively recent, have spawned a number of exciting new technological applications. These include, among many others: Automatic target detection (e.g., finding traffic signs along the road or military vehicles in a savanna; Itti & Koch, 2000); Robotics (using salient objects in the environment as navigation landmarks; Frintrop et al., 2006; Siagian & Itti, 2007); Image and video compression (e.g., giving higher quality to salient objects at the expense of degrading background clutter; Osberger & Maeder, 1998; Itti, 2004); Automatic cropping/centering of images for display on small portable screens (Le Meur et al., 2006); Medical image processing (e.g., finding tumors in mammograms; Hong & Brady, 2003); and many more.

From attention to visual scene understanding

How can cognitive control of attention be integrated to the simple framework described so far? One hypothesis is that another map exists in the brain, which encodes for top-down relevance of scene locations (a so-called Task-Relevance Map or TRM; Navalpakkam & Itti, 2005). In the TRM, a highlighted location might have a particular relevance because, for example, it is the estimated target of a flying object, it is the next logical place to look at given a sequence of action, recognition and reasoning steps, like in a puzzle video game, or one just guesses that there might be something worth looking at there. Clearly, building computational models in which a TRM is populated with sensible information is a very difficult problem, tantamount to solving most of the problems in vision and visual scene understanding.

Recent work has demonstrated how one can implement quite simple algorithms which deliver highly simplified but useful task-relevance maps. In particular, Peters & Itti (2007; Figure 3) recently proposed a model of spatial attention that (1) can be applied to arbitrary static and dynamic image sequences with interactive tasks and (2) combines a general computational implementation of both bottom-up (BU) saliency and dynamic top-down (TD) task relevance (Figure 5). The novelty lies in the combination of these elements and in the fully automated nature of the model. The BU component computes a saliency map from 12 low-level multi-scale visual features, similar to described in the previous section. The TD component computes a low-level signature of the entire image (the so-called "gist" of the scene; Torralba, 2003; Peters & Itti, 2007), and learns to associate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest. It is important to note that while we call this

component top-down, that does not necessarily imply that it is high-level --- in fact, while the learning phase certainly uses high-level information about the scenes, that becomes summarized into the learned associations between scene signatures and expected behavior in the form of TD gaze-position maps.

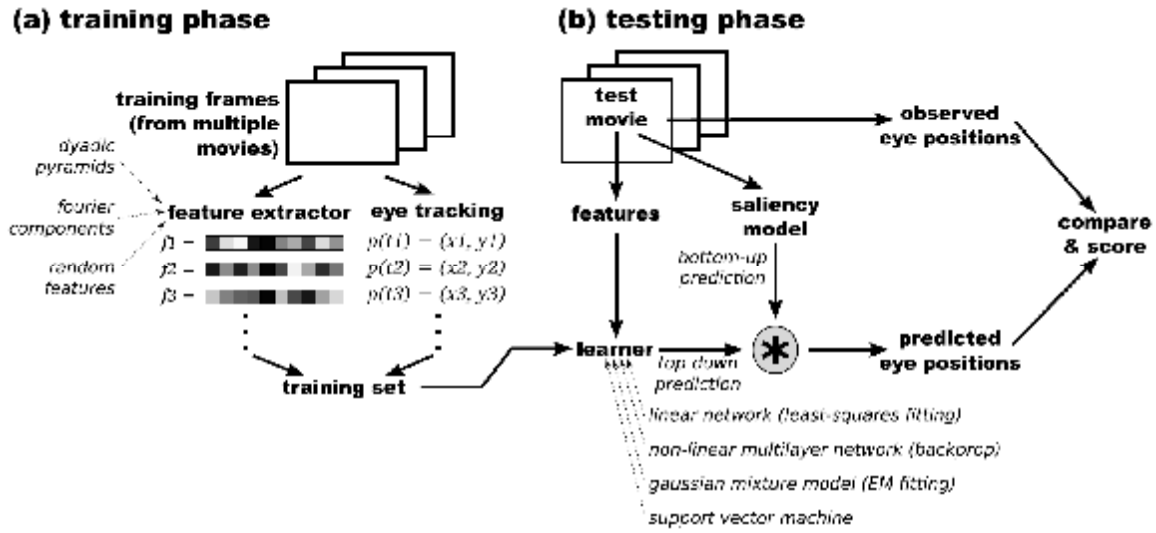

Figure 3: Schematic illustration of our model for learning task-dependent, top-down influences on eye position. First, in (a) the training phase, we compile a training set containing feature vectors and eye positions corresponding to individual frames from several video game clips which were recorded while observers interactively played the games. The feature vectors may be derived from either: the Fourier transform of the image luminance; or dyadic pyramids for luminance, color, and orientation; or as a control condition, a random distribution. The training set is then passed to a machine learning algorithm to learn a mapping between feature vectors and eye positions. Then, in (b) the testing phase, we use a different video game clip to test the model. Frames from the test clip are passed in parallel to a bottom-up saliency model, as well as to the top-down feature extractor, which generates a feature vector that is used to generate a top-

down eye position prediction map. Finally, the bottom-up and top-down prediction maps can be combined via point-wise multiplication, and the individual and combined maps can be compared against the actual observed eye position.

We measured (Peters & Itti, 2007) the ability of this model to predict the eye movements of people playing contemporary video games (Figure 4). We found that the TD model alone predicts where humans look about twice as well as does the BU model alone; in addition, a combined BU*TD model performs significantly better than either individual component. Qualitatively, the combined model predicts some easy-to-describe but hard-to-compute aspects of attentional selection, such as shifting attention leftward when approaching a left turn along a racing track. Thus, our study demonstrates the advantages of integrating bottom-up factors derived from a saliency map and top-down factors learned from image and task contexts in predicting where humans look while performing complex visually-guided behavior. In continuing work we are exploring ways of introducing additional domain knowledge into the top-down component of our attentional system.

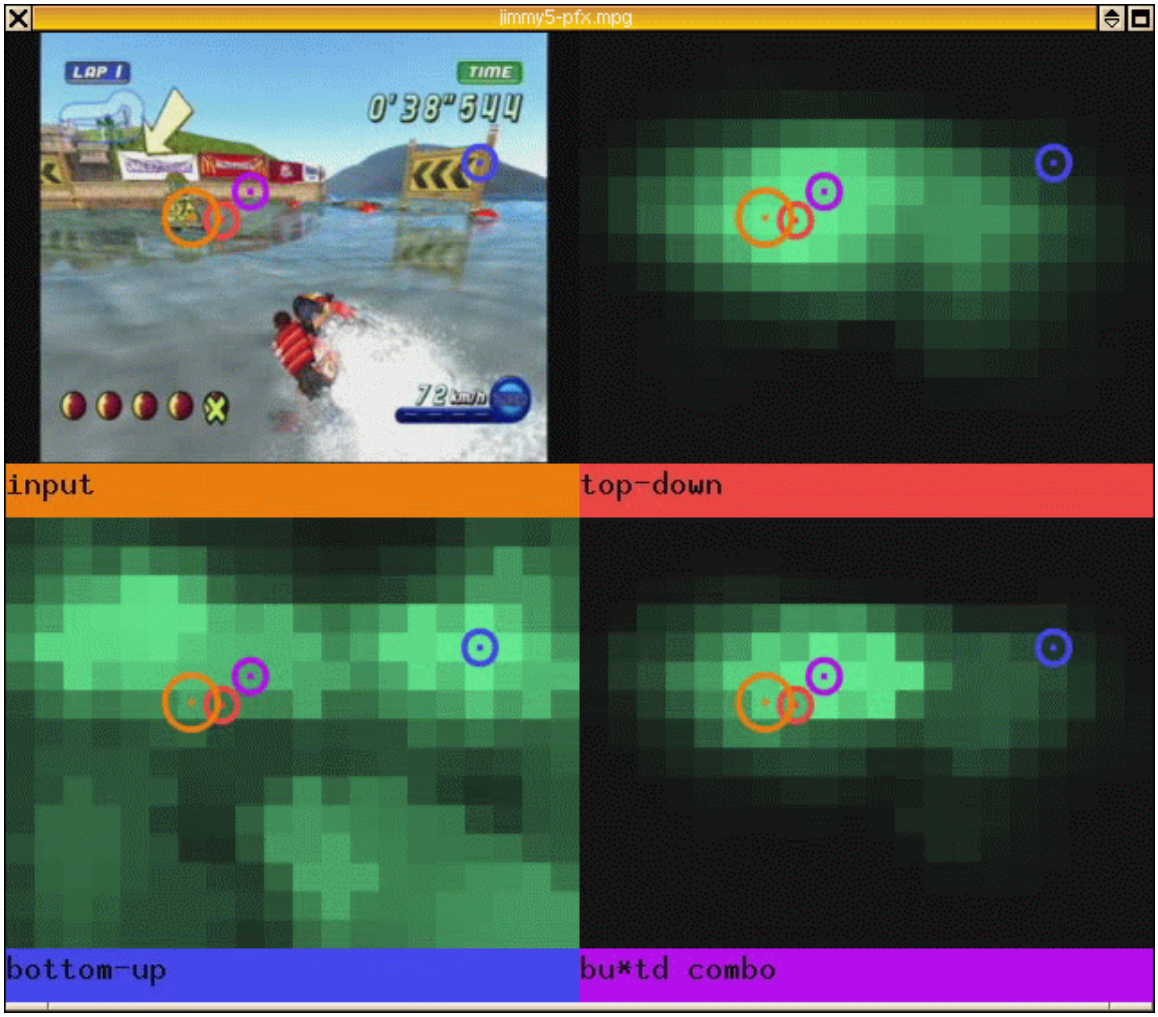

Figure 4: Combined bottom-up + top-down model and comparison with human eye movements (Peters & Itti, 2007). While a subject was playing this jet-ski racing game (upper-left), his gaze was recorded (large orange circle in all four quadrants). Simultaneously, the top-down map (upper-right) was computed from previously learned associations between scene gist and human gaze (location of maximum top-down activity indicated by small red circle). The bottom-up saliency map was also computed from the video input (lower-left, maximum at small blue circle). Finally, both maps where combined, here by taking a pointwise product (lower-right, maximum at small purple

circle). This figure exemplifies a situation where top-down dominates gaze allocation (several objects in the scene are more bottom-up salient than the one being looked at).

Discussion and conclusions

Visual processing of complex natural environments requires animals to combine, in a highly dynamic and adaptive manner, sensory signals that originate from the environment (bottom-up) with behavioral goals and priorities dictated by the task at hand (top-down). In the visual domain, bottom-up and top-down guidance of attention towards salient or behaviorally relevant targets have both been studied and modeled extensively, a well as, more recently, the interaction between bottom-up and top-down control of attention. In recent years, thus, a number of neurally-inspired computational models have emerged which demonstrate unparalleled performance, flexibility, and adaptability in coping with real-world inputs. In the visual domain, in particular, such models are achieving great strides in tasks including focusing attention onto the most important locations in a scene, recognizing attended objects, computing contextual information in the form of the "gist" of the scene, and planning/executing visually-guided motor actions, among many other functions.

Together, these models present great promise for future integration with more conventional Artificial Intelligence techniques: Symbolic models from artificial intelligence have reached significant maturity in their cognitive reasoning and top-down abilities, but the worlds in which they can operate have been necessarily simplified (e.g., a chess board, a virtual maze). Very exciting opportunities exist at present to attempt to bridge the gap between the two disciplines of neural modeling and artificial intelligence.

Acknowledgements: This work is supported by HFSP, NSF, NGA and DARPA.

References:

J. Allman, F. Miezin, & E. McGuinness (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. Annual Review of Neuroscience 8:407-430.

N. Bruce & J. K. Tsotsos (2006). Saliency Based on Information Maximization. In: Advances in Neural Information Processing Systems, 18:155-162.

M. W. Cannon & S. C. Fullenkamp (1991). Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. Vision Research 31:1985-1998.

C. L. Colby & M. E. Goldberg (1999). Space and attention in parietal cortex. Annu. Rev. Neurosci. 22:319-349.

F. Crick (1984). Function of the thalamic reticular complex: the searchlight hypothesis. Proceedings of the National Academies of Sciences USA 81(14):4586-90.

R. Desimone & J. Duncan (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18:193-222.

S. Frintrop, P. Jensfelt, & H. Christensen (2006). Attentional Landmark Selection for Visual SLAM. In: Proc. IEEE International Conference on Intelligent Robots and Systems (IROS'06).

J. P. Gottlieb, M. Kusunoki & M. E. Goldberg (1998). The representation of visual salience in monkey parietal cortex. Nature 391:481-484 (1998).

B.-W. Hong & M. Brady (2003). A Topographic Representation for Mammogram Segmentation. In: Lecture Notes in Computer Science 2879:730-737.

D. H. Hubel & T. N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol (London), 160, 106-54.

L. Itti, C. Koch, & E. Niebur (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11):1254-1259.

L. Itti & C. Koch (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40(10-12):1489-1506.

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

L. Itti (2004). Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. IEEE Transactions on Image Processing 13(10):1304-1318.

L. Itti & P. Baldi (2006). Bayesian Surprise Attracts Human Attention. In: Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005), Cambridge, MA:MIT Press.

C. Koch & S. Ullman (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4:219-227.

S. W. Kuffler (1953). Discharge patterns and functional organization of mammalian retina. J. Neurophysiology, 16:37-68.

O. Le Meur, P. Le Callet, D. Barba, & D. Thoreau (2006). A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(5):802-817.

M. Minsky (1961). Steps Toward Artificial Intelligence. In Proceedings of the Institute of Radio Engineers 49:8-30. New York: Institute of Radio Engineers.

V. Navalpakkam & L. Itti (2005). Modeling the influence of task on attention, Vision Research 45(2):205-231.

A. Newell and H. Simon (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

E. Niebur & C. Koch (1996). Control of Selective Visual Attention: Modeling the `Where' Pathway. Neural Information Processing Systems 8:802-808.

W. Osberger & A. J. Maeder (1998). Automatic Identification of Perceptually Important Regions in an Image using a Model of the Human Visual System. International Conference on Pattern Recognition, Brisbane, Australia.

D. L. Robinson & S. E. Petersen (1992). The pulvinar and visual salience. Trends Neurosci. 15:127-132.

R. Rosenholtz (1999). A simple saliency model predicts a number of motion popout phenomena. Vision Research, 39:3157-3163.

A. Samuel (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal 3(3):210-229.

C. Siagian & L. Itti (2007). Biologically-Inspired Robotics Vision Monte-Carlo Localization in the Outdoor Environment, In: Proc. IEEE International Conference on Intelligent Robots and Systems (IROS'07).

A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, & J. Davis (1995). Visual cortical mechanisms detecting focal orientation discontinuities. Nature 378:492-496.

A. Torralba (2003). Modeling global scene factors in attention. Journal of the Optical Society of America A - Optics Image Science and Vision, 20(7):1407-1418.

A. Treisman G. & Gelade (1980). A feature integration theory of attention. Cognitive Psychology 12:97-136.

A. Treisman (1988). Features and objects: The fourteenth Bartlett memorial lecture. Q. J. Exp. Psychol. A 40:201-237.

J. K. Tsotsos (1991). Is Complexity Theory appropriate for analysing biological systems? Behavioral and Brain Sciences 14(4):770-773.

E. Weichselgartner & G. Sperling (1987). Dynamics of automatic and controlled visual attention. Science 238:778-780.

J. M. Wolfe (1994). Guided Search 2.0: A Revised Model of Visual Search. Psychonomic Bulletin & Review 1(2):202-238.

J. M. Wolfe (1998). Visual Search. In: Pashler H., editor. Attention. London UK: University College London Press.

J. M. Wolfe & T. S. Horowitz (2004). What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5:1-7.