Lecture notes on visual attention

R. Brémond

Université Paris Est, LEPSiS, INRETS-LCPC

Abstract

Abstracts of reference papers in the field of visual attention and visual saliency are proposed. Personal remarks are sometimes included.

Keywords: visual attention, visual saliency, computational biology, visual search, pop-out

1. A brief and selective history of attention

(Tsotsos et al., 2005)

Descartes' hypothesis that attention is controlled by the pineal body (Descartes, 1649) is not considered any more[1]. Many alternative models have been proposed since, both theoretical and computational, to describe what appeared to James (1890) as an obvious concept: "Everyone knows what attention is."

The first psychophysical studies on attention investigated the span of visual attention, and von Helmoltz (1896/1989) soon proposed the concept of covert attention: the deployment of visual attention may be independent of eye movements. While early approaches considered attention as a top-down guided internal state, the Gestalt-theory tend to consider the focus of attention as a bottom-up computation (Köhler, 1947).

Sechenov's neuro-physiological findings that inhibition from the central nervous system may be a key process in attention control (Sechenov, 1863/1965) gave rise, more than one century later, to a number of computational models of attention, including Tsotsos' one (Tsotsos et al., 1995). Inhibition was considered as a theoretical level by Pavlov as an attention process (Pavlov, 1927): an unexpected stimulus tends to capture attention (facilitation) (see also Itti and Baldi (2009)), while stimuli in the near cortical areas are faded by an inhibitory process.

Cognitive psychology proposed the first comprehensive model of attention (Broadbent, 1958), described as a biological trick to cope with the limited capacity of the information process. This bottom-up theory of attention raised a discussion about the stage where the selection happens: early in the information pipeline (Broadbent, 1958), or latter in the selection process ? (Norman, 1968). Treisman proposed a half-way between these two options, including an attenuation of unattended signals (Treisman, 1964). Shiffrin and Schneider (1977) split the information processes between serial processes, which are conscious, slow and limited in resources, and parallel processes, which are unconscious, fast and virtually unlimited in resources. One interesting insight of their work is to propose that practice may move a task from the serial to the parallel process.

Milner (1974) was among the first to propose that the attention not only selects relevant features, but also send feedback to the early stages of the information processing. This framework was then implemented in Grossberg's Adaptive Resonance Theory (Grossberg, 1975). Evidence from neuro-physiology showed since that the attention state impacts, in a

[1] As far as I know, the **International Journal of Paper Abstracts** does not exist.

top-down manner, the activation state of the perceptual circuitry, and it became clear that such feedback may happen at any stage of the information pipeline.

Inhibition Of Return (IOR) refers to a bias against attention focus on areas previously attended, which may contribute to optimize the visual search sampling (Posner et al., 1985).

Most computational models of visual saliency originates from the Feature Integration Theory (FIT) of spatial visual attention (Treisman and Gelade, 1980). The main purpose of this model was to explain the difference in performance between pop-out stimuli and conjunction search. Bergen and Julesz (1983) also showed in their *texton* theory that some features allow fast discrimination between a target and surrounding outliers (pop-out), while other don't (in the latter case, the discrimination time is a function of the number of outliers). According to the FIT, a unique saliency map ("master map") gathers informations from separate feature maps about salient locations. Treisman and Gelade also addressed the binding issue (Rosenblatt, 1961): the unified representation of an object implies that the object's features (color, shape, location, etc.) are bind together in some way. Koch and Ullman (1985) proposed since a computational implementation of the FIT, where the saliency map is a weighted sum of the feature maps. A Winner-Takes-All competition between the salient regions leads to the selection of the current focus of attention, and the IOR is implemented in order to select the next salient location.

Although most models of visual attention are models of covert attention, eye tracking studies have extensively addressed the link between covert and overt attention, through eye movements, showing a strong top-down dependence (Yarbus, 1967). Posner (1980) linked overt and covert attention into a unique framework, through the three functions devoted to the attentional system: *alerting*, *orienting* and *search*.

2. Shifts in selective visual attention: towards the underlying neural circuitry

(Koch and Ullman, 1985)

Koch & Ullman propose a biologically plausible model for the shift of selective visual attention. Elementary features, such as color, orientation, direction of movement, disparity, etc. are coded in topographical maps. A central representation, which is not topographic, contains the properties of the selected focus of attention. The major rules, in order to select the focus of attention from the feature maps, is implemented using a WTA network. Inhibiting the current focus of attention (Shiffrin and Schneider, 1977) leads to a shift towards the next most salient location. Other rules are discussed, such as *proximity* and *similarity*.

The standard paradigm for detection, localization and recognition of objects includes two steps: a preattentive one, where the entire visual field is processed in parallel, and an attentive one, which processes the information in the focus of attention. high level processing is associated with the second step, called *selective attention*.

Treisman and Gelade (1980) showed that visual search for targets defined by a single feature (e.g. red vs. green) occurs in parallel (pop-out), whereas targets defined by the conjunction of several features (e.g. a red horizontal bar) requires serial processing, scanning the distractors in the visual field. This results in constant search latency (vs. number of distractors) in single feature search, and linear search latency in conjunction search. Similar results were shown by Julesz (1984) about texture discrimination: only a limited set of texture features (*textons*) are detected in parallel. Reported elementary features are color, orientation of line segments and curvature.

The questions that arise are: what operations apply to the selected location? how is this location selected?

The problem. Koch and Ullman (1985) suggest that selective visual attention operates on topographic cortical maps called *early representation*, coding for various elementary features such as orientation, color, etc. Local connections in these maps (or before) implement lateral inhibition, which result in contrast feature selection: areas where the feature itself is constant are not selected, whatever the absolute value. These maps may well exist at different spatial scale

(Campbell and Robson, 1968); they code feature conspicuity. Then, the area in the focus of attention is processed in a non-topographic (conscious ?) representation.

From this framework, two problems arise: (1) the spatial selection should select only one area at a time. (2) spatial accuracy: how biological system may keep the topographic information across the various feature maps, in order to select the focus of attention? (3) how does it work to shift from a location to another? Two biologically plausible mechanisms are proposed, computing the feature conspcicuity at a given location (saliency map) and selecting the most active unit in such a map (WTA).

The saliency map. The feature maps code for the conspicuity within a feature dimension. The (hypothetical) Saliency Map (SM) combines information from the conspicuity maps into a global measure of conspicuity. Saliency rates how different a location is compared to its surround. Of course, top-down modulation is possible.

Winner Takes All. The next step is the attention selection in the saliency map. Two implementation are proposed for a WTA network, which selects the most active unit in the SM (biological *maximum*). Then, a second network directs the properties of the selected area in the central representation.

The simplest WTA implementation is a mutual inhibitory network: every unit inhibit every other unit (Hadeler, 1974). In the end, only the higher units are non-zero. However, this model does not converge, and needs many connexions. An alternative model may be derived from Hadeler's equation:

$$\frac{\partial y_i}{\partial t} = y_i(x_i - \Sigma_j x_j y_j) \tag{1}$$

where x_i denotes the SM, and y_i the WTA map. The static solution is (if x_i is constant over time):

$$y_i = \frac{y_i(0)e^{x_i t}}{\Sigma_j y_j(0)e^{x_j t}} \tag{2}$$

As $\forall t, \Sigma_j y_j = 1$, y may be seen as a probability distribution. Note that the convergence speed depends on the SM activity. Computationally, only one fast computing unit is needed, in order to compute $\Sigma_j x_j y_j$, which may be seen as the network activity. If the SM changes over time, the key issue is to compare the time constants of the two processes: changes in the SM, and changes in the WTA.

The authors proposed another, faster implementation of the WTA. Two pyramidal networks are mixed. The first one selects, at each level, the stronger among the competing units (knock-out competition). However, the relevant information is not the *value* of the maximum, but its *location*. Thus, an auxiliary pyramid selects the location of the selected unit: an auxiliary unit is activated if it receives activation from both the corresponding main unit, and the auxiliary unit at the next higher level.

The last step is to copy the information under focus (the feature maps content at the selected location) into the central representation. This may be seen as a spotlight inspection (Posner, 1980). No implementation is proposed for this "copy" process.

Shifting the focus of attention. The shift of attention takes time, and the delay depends on the angular distance. Several implementations are possible, such as inhibition feedback from the central unit, or local inhibition at the SM level after time (note that both mechanisms let the feature maps unchanged). Koch and Ullman (1985) emphasize the fact that with their second WTA architecture (pyramidal), the computation of the next focus of attention is shorter when the location is closer.

The proposed mechanisms explain both parallel and serial search. When the target differs from the distracting objects by one property (a red object among green objects), its locations pops out in the feature map, and then in the SM. When a conjunction is needed, no location pops-out in the corresponding feature maps, and the focus of attention is almost randomly selected.

The proposed mechanism also explains some aspects of visual masking (camouflage). Blending an object with its background, or adding conspicuous objects around, both lower the SM activity at target location.

Shifting rules. Two additional mechanisms are proposed for the control of the attentional shift, both related to the Gestalttheorie. The *proximity* preference may be implemented by enhancing the locations close to the current focus of attention in the SM, with a factor depending on the distance. The *similarity* preference is the fact that the features contributing to the selection of the current focus of attention are enhanced when selecting the next focus. It may be implemented by increasing the conspicuity of these feature maps.

Fusion. In the central representation, the information coming from the feature maps are glued into a single object representation. When objects are *not* in the focus of attention, the features are glued on a random basis, leading to illusionary conjunctions (Treisman and Schmidt, 1982).

3. Computational modeling of visual attention

(Itti and Koch, 2001)

Most computational models of visual attention focus on the bottom-up image-based control of attention. Five trends emerge: (1) Visual saliency strongly depends on the surrounding context. (2) Using a unique saliency map efficiently models the bottom-up attention control. (3) Inhibition of return (IOR) is a key issue in attention modeling. (4) Covert attention (saliency) and overt attention (eye movements) are strongly linked, but their interaction still is a challenge for computational models. (5) Scene understanding and object recognition constrain the attention selection.

Covert attention directs the gaze toward objects of interest. The current framework is that the selection of salient items uses both bottom-up, image-based saliency cues, and top-down task-dependent cues. The bottom-up selection is massively parallel, involuntary, rapid and automatic (Shiffrin and Schneider, 1977); however it is not completely straightforward, in the sense that the local surround strongly modulates this selection, through center-surround mechanisms at different spatial scales. Top-down attention, in contrast, is costly in cognitive resources.

The behaviorally relevant part of visual information is selected and reaches the short term memory (Sperling, 1960). In Broadbent's framework, it is a way to cope with the limited processing capacity of the nervous system (Broadbent, 1958).

Binding is a key function of top-down feedback, and overall of conscious representation. Attention not only selects a region of interest (the *where* (dorsal) visual stream), it also enhances the object's representation (the *what* (ventral) visual stream).

The aim of this review is biologically plausible computational models of bottom-up (covert) attention. The key concept of most of these models is the saliency map (Koch and Ullman, 1985), and thus originates from Tresiman's Feature Integration Theory (FIT) (Treisman and Gelade, 1980). Top down attention is not reviewed, due to the lack of computational models.

Pre-attentive computation of visual features. Early visual features are computed in the first processing stages of the visual pipeline. They may code intensity contrast, color oponency, orientation, direction and velocity of motion, stereo disparity, etc. (Wolfe and Horowitz, 2004), at several spatial scales.

We are far less sensitive to what happens outside the focus of attention, which may be set in terms of higher psychophysical thresholds. Models have been proposed in terms of enhanced gain, biased competition, intensified competition (Winner Takes All[2]), enhanced spatial resolution, modulated background activity, stimulus strength and noise, etc.

The early stages of visual processing are usually described in terms of center-surround filters (Difference of Gaussian, DoG) (Wandell, 1995). Similarly, orientation selection is modeled through Gabor wavelets(Daugman, 1984). The main result is that the visual system is sensitive to feature contrast rather than absolute feature levels. Interestingly, closed contours were found to be enhanced in V1.

[2]A WTA neural network is nothing more than a *maximum* detector, which links neuroscience to mathematical morphology (Serra, 1982, 1988).

Once the feature maps are computed, they are weighted in the saliency map. Top-down control may change the weights (visual search, learning, etc.) (Nothdurft, 2000). One important result is the lack of interaction across features. This result is linked to the low performance in searching for conjunctive targets (Treisman and Gelade, 1980; Wolfe, 1996). There seems to be no competition either between spatial scales.

Visual Saliency. Early visual processes may be seen as a filter bank, including contextual modulation. In Koch & Ullman's model, the feature maps feed a saliency map, which maximum is seen as the focus of attention. It should be noted that the rest of the map, in this model, does not model anything. The model only predicts spatial pre-attentive selection (*where*).

An alternative model to the FIT is proposed by Wolfe et al. (1989); Wolfe (1996, 2007) for visual search. It states that parallel computation in the feature maps help the serial process to select the Focus of Attention. Interestingly, the saliency is seen as a probability: the likelihood that the target is present at a given location.

Tsotsos et al. (1995) used a hierarchy of feature extraction, bottom-up and feed-forward, followed by selective tuning feedback. The salient areas propagate back some kind of inhibition around them in terms of feature sensitivity. However, its seems that these feedback does not change the choice of the spatial focus of attention.

Milanese et al. (1994) proposed a model which is not really biologically-inspired. Relaxation optimizes "energy" in several ways: (1) biasing towards regions where several feature maps are excited; (2) grouping salient points into clusters, (3) minimizing "energy" in the feature maps; (4) maximizing the dynamic range of each map.

Itti's model (Itti et al., 1998; Itti and Koch, 2000) uses surround modulation to select salient areas in each feature map. This is implemented with a DoG in Itti and Koch (2000), followed by a *half wave rectification* (?). in order to remove non-salient areas. This model was tested in pop-out, conjonctive search tasks (Itti et al., 1998) and search asymetries (Itti and Koch, 2000). The effect of noise on the saliency map was also explored (Itti et al., 1998).

In contrast to these models, Desimone and Duncan (1995) proposes that the focus of attention may be selected without explicit saliency map. Top-down selection enhances the more relevant feature maps, and inhibit the less relevant. Hamker (1998, 2004) proposed a computational model derived from this idea. However, as in Wolfe's Guided Search, the feature biases computation needs a search task.

Inhibition of Return. The saliency map tries to model covert attention: where the focus of attention will shift next? However, covert and overt attention (eye movements) are strongly linked.

The scan-path is computed in (Koch and Ullman, 1985) from the saliency map with successive inhibitions of the most salient areas. This is a simulation of the biological Inhibition of Return (Klein, 2000), however with important differences. For instance, biological IOR is object based, and follows moving objects.

Attention and Recognition. The previous models may well predict the visual behavior in the second after the presentation of a new scene. To ask for more, a mixed model, including top-down and bottom-up processes as well as overt and covert attention is needed.

Several models try to predict the visual scan-path. Schill et al. (2001) proposes that the focus of attention selects informative areas in order to lower ambiguity. Ryback et al. (1998) also proposes a model with both bottom-up and top-down processes. Bottom-up selects the *where*, while top-down selects the *what* of the *where*. Deco and Schumann's model first selects a set of *where* at a coarse scale, then these locations are processed in a *what* manner (object recognition) at finer and finer scales, until an object is found Deco and Schumann (2000). Stark et al. (2001) states that the control of eye movements is mostly top-down (*scan-path theory*). The cognitive model of what we expect is the basis of our percepts. Note that all these complex models of the visual behavior fail in finding a biologically plausible framework.

4. A model of saliency-based visual attention for rapid scene analysis

(Itti et al., 1998)

A computational model of visual attention is proposed, based on a biologically plausible architecture (Koch and Ullman, 1985) related to the FIT (Treisman and Gelade, 1980). Other architectures have been proposed, such as *dynamic routing* (Olshausen et al., 1993). Koch and Ullman's model was also used by previous authors (Milanese et al., 1995; Baluja and Pomerleau, 1997). This model represents the bottom-up saliency, and does not requires any top-down guidance to shift attention, in contrast with Wolfe (1994), which is a model of visual search, including a representation of the target to enhance the relevant features (see also Gao and Vasconcelos (2005); Simon et al. (2008)). It is a fast parallel method for the selection of a small number of regions of interest, to be analyzed by more complex (and time-consuming) processes, such as object recognition.

Feature maps. from 640×480 images, dyadic Gaussian pyramids are built (Adelson, 1982), in order to implement the multi-scale center-surround processes in the early visual system. The Difference of Gaussian (DoG) compute center-surround operators (to compute the difference, an interpolation of the coarser scale is needed). In the following, c is the scale and δ is the scale difference in the DoG; the across-scale difference is denoted $\ominus$. The authors use $c \in \{2,3,4\}$ and $\delta \in \{3,4\}$.

An *intensity* image is first obtained from $I = (R + G + B)/3$, and an *intensity* pyramid is built from. The r, g and b channels are normalized by I, except that they are set to 0 when $I \leq I_{max}/10$. Four color channels are created, leading to four color pyramids:

- $R^\star = max(0, r - (g + b)/2)$ (red)
- $G^\star = max(0, g - (r + b)/2$ (green)
- $B^\star = max(0, b - (r + g)/2$ (blue)
- $Y^\star = max(0, (r + g)/2 + |r - g|/2 - b)$ (yellow)

Center-surround differences are computed on the intensity channel: $I(c, s) = |I(c) \ominus I(s)|$. Thus, the *ON* and *OFF* channels of biological vision are mixed.

Color features are computed using red/green and blue/yellow opponencies (Gegenfurtner and Kiper, 2003). Two channels are computed:

- $RG^\star(c, s) = |(R^\star(c) - G^\star(c)) \ominus (G^\star(s) - R^\star(s))|$
- $RG^\star(c, s) = |(B^\star(c) - Y^\star(c)) \ominus (B^\star(s) - Y^\star(s))|$

Note that most of the time (when the max is > 0), $R^\star - R^\star = 3/2(r - g)$, and $B^\star - Y^\star = 2b - 3g/2 - r/2$ (or $2b - 3r/2 - g/2$, depending on the sign of $r - g$).

Orientation features are computed using oriented Gabor pyramids, based on the *intensity* channel. Then,

- $O^\star(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$

Finally, 42 feature maps are computed: 6 for intensity, 12 for color and 24 for orientation ($\theta = 0°, 45°, 90°$ and $135°$).

Saliency map. To combine the feature maps in a single saliency map, Itti et al. (1998) propose a 3-steps normalization operator **N**, to enhance rare strong peaks, and suppress noise. The first step is to normalize the current map to a fixed range (say [0...1]). Then, the average $\bar{m}$ of all local maxima $m_i < 1$ is computed. Finally, the map is multiplied by $(1 - \bar{m})^2$. The authors claim that this mechanism coarsely replicates lateral inhibition (Cannon and Fullenkamp, 1996).

The normalized maps are summed up across scale into global maps: one for intensity, two for color and four for orientation. The color maps ($RG^\star$ and $BY^\star$) are added in a unique color map, and the orientation maps are normalized again into a unique orientation map. Finally, the 3 intermediate conspicuity maps are normalized, then summed up in the Saliency Map (SM).

Focus of Attention. The Focus of Attention (FOA) is the maximum of the SM. It is implemented as a 2D layer of *integrate and fire* neurons $SM^\star$, at scale 4 of the image pyramid, fed by the SM and feeding a WTA network. the SM pixels send a continuous signal to

$SM^\star$ neurons, proportional to their activity. When a threshold is reached, the $SM^\star$ neurons fire to the WTA, and are reseted to 0. In parallel, the WTA computes the most active location (the FOA). When a new FOA is selected, all WTA neurons are reseted, and a local inhibition is activated in the SM around the current FOA (inhibition of return).

The FOA was implemented as a disc of radius 80 pixels. Time constants were set so as to jump from one location to the next one in 30-70 ms, and the inhibition time last 500-900 ms.

Results. The model was compared to another approach: rating the entropy (Spatial Frequency Content, SFC) as a measure of visual saliency. The proposed approach seemed very robust to additional noise, while the SFC approach wasn't.

The more interesting result is the reproduction of pop-out for laboratory tasks, while conjunctive search needed search time which linearly increased with the number of distractors. In natural images, the SM was quite similar to the SFC, except in some extended areas with high SFC and low saliency, where the saliency map seemed more relevant.

5. Guided Search: an alternative to the Feature Integration model for visual search

(Wolfe et al., 1989)

Searching for a target among distractors is easier for some stimuli than for others. When the target differs from the distractors by a unique feature, the Reaction Time (RT) is almost independent from the number of distractors (Treisman and Gelade, 1980), while for conjunctive searches, RT linearly depends on the number of distractors (e.g. a **T** among **L**s). Moreover, for "serial" search, the RT when a target is present is expected to be half the RT when no target is present.

Treisman's FIT is the main model to explain the difference between serial and parallel visual search (Treisman and Gelade, 1980; Treisman, 1986). Two steps are proposed: a pre-attentive, massively parallel one computes basic feature maps, and a serial one is needed for conjunction search. Julesz's *texton* model (Julesz, 1984) shares many features with Treisman's one.

One problem with these models is that the parallel computation has little influence on the serial search process. Wolfe et al. (1989) conducted several visual search experiments, suggesting that visual search may be guided by informations from the parallel processes.

A first series of experiments resulted in conjunctive search (color + form, color + orientation) with very flat slopes in the "RT vs. set size" function. The difference with Treisman's results was investigated in a second series of experiments. The conclusion was that some changes are needed in the FIT. Wolfe proposed that the parallel process *guides* the spotlight of attention towards likely targets. This model is consistent with Hoffman (1979).

Unlike Treisman's model, the guided search predicts that triple conjunctions are easier to detect than double conjunction, that is, lead to flatter slopes in the RT vs. set size function, which was found experimentally.

Experiments 1-3: conjunction search tasks. Three conjunction search tasks were tested: Color × Form, Color × Orientation and Color × Size. In the first one for instance, targets were red **O**s, while distractors were green **O**s and red **X**s. In all three experiments, the slopes of RT vs. set size were much lower than the Treisman's experiment (Treisman and Gelade, 1980). In addition, RT vs. set size data were not linear in the Color × Orientation task (may be due to side effects, such as density and image border). The Color × Size experiment was consistent with Treisman and Gormican (1988).

*Experiment 4: **T** vs. **L**s.* In contrast with these conjunction search experiments, a search experiment used **T** as target and **L**s as distractors, both with 0°, 90°, 180° or 270° rotations, with results consistent with the Feature Integration Theory. Wolfe's understanding is that the feature maps, in this case, do not covey relevant information to the serial process.

Experiment 5 and 6: practice effects. Two experiments were conducted in order to control possible

practice effects. In one case (conjunctive search), an effect of practice was found for blank trials, not for target trials. Conversely, for **T** vs. **L** searches, a practice effect was found for target trials, not for blank trials. However, one may argue that a different statistical analysis ma have found stronger practice effects in all situations.

Experiment 7 and 8: stimulus salience. Treisman and Gormican (1988) argued that less salient targets and distractors lead to stepper slopes in the RT vs. set size graphs, which was also found in Wolfe's experiments (under mesopic light levels), however with slopes lower thane expected. Anyhow, it appears that the target and distractor's saliency plays a role in determining whether the search for conjunction is serial or parallel, which enforces the Guided Search hypothesis.

The Guided Search model. Wolfe's understanding of the data was that the FOA can be guided by preattentive, parallel mechanisms, which select candidate targets. Thus, eye movements are not random, they are directed towards the most likely target. If the signal from the parallel process is high (compared to noise), the target is found quickly. Note that in Guided Search, the difference between serial and parallel search is only quantitative. Wolfe notes that even in Treisman's approach, the parallel process selects some information: fixations do not explore blank areas!

Guided Search is implemented with an addition of the feature map selection (with reference to the known target), followed by the selection of the maximum value of this target-based saliency map. The psychological meaning is that some top-down process (selection of the relevant features) reaches the parallel process. Visual performance is explained by noise in the parallel $\mapsto$ serial information transmission. The parallel process follows on during all the stimulus duration, so that threshold may be overtaken during visual search.

The Guided Search explains the difference between conjunctive search and **T** vs. **L** searches. It also predicts that conjunctive search with 3 features are easier than with 2 features, which was demonstrated in an additional experiment.

References

Adelson, E. H., 1982. Saturation and adaptation in the rod system. Vision Research 22, 1299–1312.

Baluja, S., Pomerleau, D. A., 1997. Expectation-based selective attention for visual monitoring and control of a robot vehicle. Robotics and Autonomous systems 22 (3-4), 329–344.

Bergen, J. R., Julesz, B., 1983. Parallel versus serial processing in rapide pattern discrimination. Nature 303, 696–698.

Broadbent, D., 1958. Perception and Communication. Pergamon Press, New York.

Campbell, F. W., Robson, J. G., 1968. Application of fourier analysis to the visibility of gratings. Journal of Physiology 197, 551–566.

Cannon, M. W., Fullenkamp, S. C., 1996. A model for inhibitory lateral interaction effects in perceived contrast. Vision Research 36 (8), 115–125.

Daugman, J. G., 1984. Spatial visual channels through the Fourier plane. Vision Research 24, 891–910.

Deco, G., Schumann, A. B., 2000. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. Vision Research 40, 2845–2859.

Descartes, R., 1649. Les passions de l'âme. Le Gras, Paris.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review Neuroscience 18, 193–222.

Gao, D., Vasconcelos, N., 2005. Discriminant saliency for visual recognition from cluttered scenes. In: Saul, L. K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems. Vol. 17. MIT Press, pp. 481–488.

Gegenfurtner, K. R., Kiper, D. C., 2003. Color vision. Annual Review Neuroscience 26, 181–206.

Grossberg, S., 1975. A neural model of attention, reinforcement, and discrimination learning. International Review of Neurobiology 18, 263–327.

Hadeler, K. P., 1974. On the theory of lateral inhibition. Kybernetik 14, 161–165.

Hamker, F., September 1998. The role of feedback connections in task-driven visual search. In: Proc. of the 5th Neural Computation and Psychology Workshop. Springer Verlag, University of Birmingham, England, pp. 252–261.

Hamker, F. H., 2004. A dynamic model of how feature cues guide spatial attention. Vision Research 44, 501–521.

Hoffman, J. E., 1979. A two stage model of visual search. Perception and Psychophysics 25, 319–327.

Itti, L., Baldi, P. F., May 2009. Bayesian surprise attracts human attention. Vision Research 49 (10), 1295–1306.

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

Itti, L., Koch, C., 2001. Computational modeling of visual attention. Nature Reviews Neuroscience 2 (3), 194–203.

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

James, W., 1890. Principles of psychology. Holt, New York.

Julesz, B., 1984. A brief outline of the texton theory of human vision. Trends in Neuroscience 7, 41–48.

Klein, R. M., 2000. Inhibition of return. Trends in Cognitive Science 4, 138–147.

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

Köhler, W., 1947. Gestalt psychology. Livernight, New York.

Milanese, R., Gil, S., Pun, T., 1995. attentive mechanisms for dynamic and static scene analysis. Optical Engineering 34 (8), 428–434.

Milanese, R., Weschler, H., Gil, S., Bost, J., Pun, T., 1994. Integration of bottom-up and top-down cues for visual attention using non linear relaxation. In: Proceedings of CVPR. IEEE, pp. 781–785.

Milner, P., 1974. A model for visual shape recognition. Psychological review 81, 521–535.

Norman, D., 1968. Towards a theory of memory and attention. Psychological review 75, 522–536.

Nothdurft, H., 2000. Salience from feature contrast: additivity across dimensions. Vision Research 40, 1183–1201.

Olshausen, B. A., Anderson, C. H., Essen, D. C. V., 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neuroscience 4, 700–719.

Pavlov, I. P., 1927. Conditioned reflexes. Oxford University Press, London.

Posner, M. I., 1980. Orienting of attention. Quarterly Journal of Experimental Psychology 32 (1), 3–25.

Posner, M. I., Rafal, R. D., Choate, L. S., Vaughan, J., 1985. Inhibition of return: neural basis and function. Cognitive neuropsychology 2 (3), 211–228.

Rosenblatt, F., 1961. Principles of neurodynamics: perception and the theory of brain mechanisms. Spartan Books, Washington DC.

Ryback, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., Shevtsova, N. A., 1998. A model of attention-guided visual perception and visual recognition. Vision Research 38, 2387–2400.

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetsche, C., 2001. Scene analysis with saccadic eye movements: top-down and bottom-up modelling. Journal of Electronic Imaging 10 (1), 152–160.

Sechenov, I. M., 1863/1965. Reflexes of the brain (trad. S. Belsky). MIT Press, Cambridge, Mass.

Serra, J., 1982. Image Analysis and Mathematical Morphology, Vol. I. Academic Press, London.

Serra, J., 1988. Image Analysis and Mathematical Morphology, Vol. II: Theoretical Advances. Academic Press, London.

Shiffrin, R. M., Schneider, W., 1977. Controlled and automatic human information processing ii: Perceptual learning, automatic attending, and a general theory. Psychological Review 84, 127–190.

Simon, L., Tarel, J.-P., Brmond, R., January 2008. Towards the estimation of conspicuity with visual priors. In: Proc. VISAPP. Vol. 2. Madère (Portugal), pp. 323–328.

Sperling, G., 1960. The information available in brief visual presentations. Psychological monographs: general and applied 74 (498), 1–21.

Stark, L. W., Privitera, C. M., Yang, H., Azzariti, M., Ho, Y. F., Blackmon, T., Chernyak, D., 2001. Representation of human vision in the brain: How does human perception recognize images? Journal of Electronic Imaging 10 (1), 123–151.

Treisman, A., 1964. The effect of irrelevant material on the efficiency of selective listening. American Journal of Psychology 77, 533–546.

Treisman, A., 1986. Features and objects in visual processing. Scientific American 255 (11), 114–125.

Treisman, A., Gormican, S., 1988. Feature analysis in early vision: evidence from search asymetries. Psychological Review 95.

Treisman, A., Schmidt, H., 1982. Illusionary conjunctions in the perception of objects. Cognitive Psychology 14, 107–141.

Treisman, A. M., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12, 97–136.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., Nuflo, F., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78 (1-2), 507–545.

Tsotsos, J. K., Itti, L., Rees, G., 2005. A brief and selective history of attention. In: Itti, L., Rees, G., Tsotsos, J. K. (Eds.), Neurobiology of attention. Elsevier, pp. xxiii–xxxii.

von Helmoltz, H., 1896/1989. Physiological optics (trans. M. Mackeben). Vision Research 29 (11), 1631–1647.

Wandell, B., 1995. Foundations of vision. Sinauer associates, Sunderland, MA, USA.

Wolfe, J. ., Cave, K. R., Fraenzel, S. L., 1989. Guided search: an alternative to the feature integration model for visual search. Journal of experimental psychology: human perception and performance 15 (3), 419–433.

Wolfe, J. M., 1994. Guided search 2.0: a revised model of visual serach. Psychonomic Bulletin Review 1 (2), 202–238.

Wolfe, J. M., 1996. Visual search. In: Pashler, H. (Ed.), Attention. , Psychology Press Ltd, pp. 13–74.

Wolfe, J. M., 2007. Guided Search 4.0: Current progress with a model of visual search. In: Gray, W. (Ed.), Integrated Models of Cognitive Systems. Oxford University Press, pp. 99–119.

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

Yarbus, A. L., 1967. Eye movements and vision. Plenum, New York.