PSYCHOPHYSIOLOGY

Science Review
UDC 159.9
https://doi.org/10.21702/rpj.2025.3.11

Visual Saliency: From Theoretical Assumptions to Modern High-Performance Models

Denis V. Yavna

Southern Federal University, Rostov-on-Don, Russian Federation

Corresponding author: yavna@fortran.su

Abstract

Introduction. Visual saliency refers to the perceptual quality of location in a visual scene, which manifests itself subjectively in its attractiveness to the observer and objectively in the probability of shifting attention or fixating eye movements on it. This quality arises from the integration of visual feature maps and is modulated by several central mechanisms. It is important to distinguish between the terms saliency and conspicuity; in a theoretical context, these are not the same. This review, for the first time, combines the results of computer modeling of visual saliency with a detailed discussion of the theoretical background for creating such models. The theory of feature integration proposed by A. M. Treisman is examined, along with its advantages and limitations, which provided the way for the three-level model of visual attention developed by C. Koch and S. Ullman. According to this theory, focal attention is governed by a "winner-takes-all" mechanism, which relies on a saliency map encoding the attractiveness of each fragment of the visual scene. The original theory did not describe how the saliency map is formed, and this question remains the focus of research using computer modeling. **Results and Discussion.** The results of studies on modeling visual saliency are reviewed. In particular, the early computational model by L. Itti, C. Koch, and E. Niebur, which laid the foundation for many subsequent developments, is described in detail. Approaches to modeling that preceded the advent of modern high-performance neural networks are examined, and a range of contemporary models based on deep learning technologies is presented, together with their characteristic properties. This is the first comprehensive review of saliency models published in Russian. Researchers have developed several models of practical utility, and the paper discusses their potential for real-world application.

Keywords

visual system, attention, eye movements, visual search, saliency, eye tracking, computer vision, modeling

Funding

The research was supported by the Russian Science Foundation (no. 25-18-00377)

For citation

Yavna, D. V. The influence of initial conditions on the formation of reflective loops in network thinking. Russian Psychological Journal, 22(3), 190-225. https://doi.org/10.21702/rpj.2025.3.11

Introduction

Why do we perceive the visible world around us in such a way that we often notice small details of our surroundings, but sometimes fail to see something we have been searching for unsuccessfully for a long time? And why can we periodically pay attention—or not pay attention—to the same object at different times and under different circumstances? A formal answer is that we usually pay attention to objects that have the quality of saliency. One could also answer that we notice objects that are conspicuous. Such answers may seem a little strange and could be perceived as pre-logical; however, it should be noted that conspicuousness and saliency are fairly well-formalized constructs, filled with specific content and used in a number of research areas on visual perception. Moreover, these constructs are not speculative and owe their emergence and content primarily to the experimental work of cognitive psychologists in the late 1970s and early 1980s. It is also important to note that outside of a scientific context (and in philological sciences), the word "saliency" is practically not used in the Russian language. Usually, words of the same root in Western languages, derived from the Latin salio ("jump, leap"), are translated as noticeability, significance, expressiveness, etc.; however, as special terms, these words have different meanings.

The term "visual saliency" has relatively recently entered the Russian language (Kochurko, Madani, Saburan, Golovko & Kochurko, 2015; Martynova & Balaev, 2015), is used by a fairly narrow circle of researchers, and therefore requires clarification. Visual saliency is generally understood as a property of a certain area of an image that characterizes its ability to attract the observer's attention. However, this understanding does not imply that saliency is a property inherent exclusively to the object of observation; saliency also has a subjective component.

There are two types of saliency: bottom-up and top-down. Bottom-up saliency is determined primarily by the physical properties of a fragment of the visual scene and is processed by stimulus-driven mechanisms of involuntary attention. For example, a red vertical line among many blue lines will have a high degree of bottom-up saliency (Fig. 1a) (*Strictly speaking, this example represents an extreme case where detection can theoretically be explained in terms of feature maps and saliency, without using the conceptual apparatus of saliency models; however, it illustrates well the phenomenal side of the issue under discussion)*. The perception of such stimuli is often accompanied by a pop-out effect, objectively expressed in the absence of time-consuming search costs, and subjectively in the ease and involuntariness of detection. It is important to note that early studies focused primarily on bottom-up saliency, and the term "top-down saliency" may have sounded strange in the past.

Top-down saliency is determined primarily by the perceptual task facing the observer. Such saliency is assigned to certain objects, features, or combinations thereof by the subject and is primarily addressed to the mechanisms of voluntary, goal-directed attention. For instance, a red vertical line among red horizontal and blue vertical lines (Fig. 1b) will have rather low bottom-up saliency, but if it is designated as a target in an experiment, its top-down saliency becomes significant. Saliency will increase, and in the course of sequential visual search, this line will sooner or later become the object of attention. In classical experiments with eye movement recording, the influence of the task on attention control was demonstrated by A. L. Yarbus (Yarbus, 1965). Yarbus analyzed the tracks of image viewing. Images were recorded during free viewing and specified by the instructions. His conclusion states: "The distribution of fixation points on an object, the sequence of their changes, their duration, and cyclicality are determined by the content of the object and the tasks of the observer" (Yarbus, 1965, p. 148).

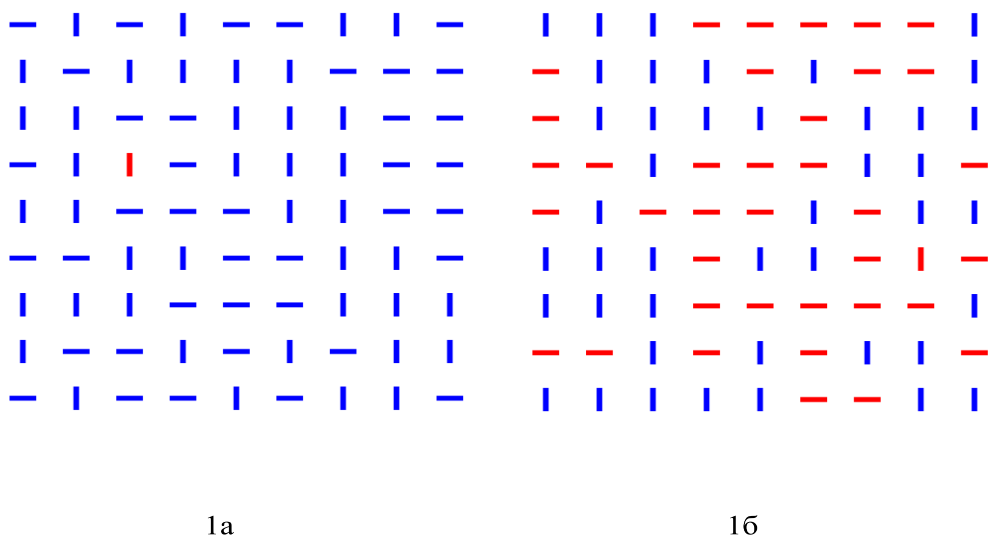
The professional experience and cultural level of the observer also have an influence. Yarbus repeatedly notes another idea-eye movements reflect the thought process. At the same time, Yarbus distinguishes between shifts in attention and eye movements. Both can be voluntary and involuntary.

Changes in the focus of attention remain in our memory, but the points of fixation are not retained.

Thus, the saliency of a particular part of an image may vary depending on the perceptual task facing the subject. Of course, the characteristics of the subject's attention also play a role in the formation of saliency, introducing additional "noise" when training computer models.

Figure 1

An example of arrays of lines in which the red vertical line is searched for in parallel (a) or sequentially (b).



Modeling visual saliency, being directly related to fundamental psychological problems such as the relationship between the focus of attention and eye movements, has a long history. While I. M. Sechenov directly identified visual attention with “the convergence of the visual axes of the eyes on the object being viewed” (Sechenov, 1942, p. 80, reprinted from the 1866 text), G. von Helmholtz (Helmholtz, 1896) demonstrated the existence of a mechanism for spatial attention shifts that does not depend on eye movements (cited in Rozhkova, Belokopytov & Iomdina, 2019). Currently, attention associated with gaze movement is commonly referred to as overt, as opposed to covert, as discovered by Helmholtz (Podladchikova et al., 2017; Rozhkova et al., 2019). Ideas about these types of attention in cognitive psychology were significantly developed by M. Posner (e.g., Posner, 1980), who later proposed a three-component model of attention (Posner & Petersen, 1990). This model is largely based on neurophysiological data and describes three subsystems of attention: alerting sistem, orienting, and executive control. The brain mechanisms and connections between implicit and explicit attention orientation are not fully understood and are a current topic of neurophysiological research (Petersen & Posner, 2012). The difficulties of objectively recording covert attention movements, combined with significant progress in eye-tracking methods, have led to the current focus on overt attention when testing models of visual saliency. Eye movements serve as an objective marker of these acts; it is believed that during fixations, the brain reads most of the information necessary for solving perceptual tasks (Rayner, 2009). Nevertheless, early work on saliency modeling focused primarily on covert attention. This apparent contradiction can be explained by the fact that both covert and overt attention operate within the same saliency map, i.e., they visit approximately the same locations, although the duration of focus and sequence of shifts may differ. Thus, “spatial attention shifts are usually (but not necessarily) accompanied by eye movements” (Theeuwes, 2013, p. 1), and eye movements “are often considered a proxy for attention shifts” (Borji & Itti, 2013, p. 186).

Theoretical Background

The theory of feature integration by A. Treisman and G. Gelade had a decisive influence on the understanding of visual saliency mechanisms. Based on early studies, the authors put forward propositions representing their theory in its “extreme form” (Treisman & Gelade, 1980, p. 99). While acknowledging that Gestalt concepts correspond to normal subjective perceptual experience, they argued that these concepts are less useful for studying early stages of information processing, where features come first. The visual scene is initially encoded according to separate features such as color, orientation, spatial frequency, brightness, and direction of movement. To synthesize these correctly for each object in a complex image, focal attention sequentially processes the corresponding locations, acting as a “glue” (Treisman & Gelade, 1980, p. 98) that connects initially separate features into a single object. Once a composite object is perceived, it is stored in memory for future recognition. Under certain circumstances (e.g., memory impairment), features may “float free” or recombine into “illusory conjunctions” (Treisman & Gelade, 1980, p. 98). Features outside the focus of attention influence task performance only at the level of individual features, not at the level of their combinations. Experiments confirmed predictions about parallel detection of basic features and the necessity of sequential scanning for conjunctions of features. Further predictions concerned figure-ground separation, illusory feature combinations, and the relationship between feature identification and localization.

The predictions made by the authors regarding various characteristics of the perception process, based on the proposed concepts, were tested in nine experiments; their results and corresponding theoretical generalizations were published in 1980 (Treisman & Gelade, 1980). Although the theory has since undergone significant development, it was this work that had the most important influence on the advancement of saliency models.

The first set of predictions stated that if basic features can be detected in parallel, without restrictions on attention, then variations in the number of simultaneously presented distractors should have little effect on the search for targets defined by such features (e.g., color red or vertical orientation). Conversely, if focal attention is required to detect targets defined by a combination of features (e.g., a red vertical line among red horizontal and blue vertical lines, Fig. 1b), such targets can only be detected after sequential scanning of the array of presented elements.

The second group of predictions dealt with the separation of textures and figure-ground grouping: if these are parallel preattentive processes, they should depend only on spatial gaps between groups of stimuli that differ in individual features, rather than in their combinations.

The third group of predictions relates to the possibility of illusory combinations of features that "float freely" outside the focus of attention.

The fourth group of predictions concerns the relationship between the identification and localization of features and their combinations. If features outside the focus of attention can float freely, and the presence of these features can be established without determining their exact location, then identification and localization are independent processes. In the case of searching for a single feature, identification may precede localization; in the case of searching for combinations, localization precedes identification, as attention is drawn to a specific location.

The fifth group of predictions pertains to the potential influence of objects outside the focus of attention on the effectiveness of the search: only features, but not their combinations, should either facilitate or hinder it.

The verification of these predictions, primarily through visual search experiments, largely confirmed their validity. B. M. Velichkovsky notes that the theory of feature integration "has withstood 20 years of experimental testing remarkably well" (Velichkovsky, 2006, p. 295), although it faced challenges in explaining the relatively flat (10–20 ms per distractor) functions relating to the dependence of search time on the number of elements. Recall that in sequential search, the slope is approximately 60 ms per element when a target is absent; if the target is present, the slope decreases by about half, indicating a potential strategy of exhaustive search: the average number of elements examined by attention before the target is found is exactly half the number of elements when the target is located in a random position. However, Velichkovsky accurately pointed out that a minimal slope in these functions would suggest viewing up to 100 elements per second, which does not align with experimental data on covert attention shifts (e.g., Saarinen & Julesz, 1991). An explanation can be offered within the framework of the theory of guided (the more common translation of the original term) or driven (according to Velichkovsky) visual search, developed by J. Wolfe et al. (Wolfe, Cave, & Franzel, 1989) /Вставить примечание: Editor's note: These are back-translations of Russian versions of the original term./. The current version of this theory (in its sixth iteration) at the time of writing this review is presented in (Wolfe, 2021).

A detailed examination of the theory of guided search is beyond the scope of this review, especially since it is well known in our country and often serves as the theoretical basis for research conducted by a number of domestic authors (e.g., Gorbunova, 2023;

Kruskop, Lunyakova, Dubrovsky & Garusev, 2023; Sapronov & Gorbunova, 2025; Falikman, 2015; Falikman, Utochkin, Markov & Tyurina, 2019)). However, it seems appropriate to give a brief summary of it in order to show the commonality of the tasks solved within the framework of the theories of guided search and saliency, as well as the similarity of their conceptual apparatus.

When we look at a scene, we can see anything in any location, but we cannot recognize more than a few elements at a time; this is a kind of bottleneck. As with Traiman, locations are selected by attention so that the features they contain can be glued together into recognizable objects. But in order for the selection order to be rational (intelligent), the attention that provides access to the "bottleneck" is guided based on five different sources of preattentive information, namely:

1. Top-down guidance
2. Bottom-up, feature-based guidance
3. Preceding history (e.g. priming)
4. Reward
5. Syntax and semantics of the scene

These sources form a spatial priority map (Serences & Yantis, 2006), a dynamic landscape of attention, with selective attention directed approximately 20 times per second (every 50 ms) to the most active location. The nature of foveal bias toward locations near the fixation point is described by three types of functional visual fields (FVF): resolution FVF, exploratory eye movement control, and covert attention control. Looking ahead a little, we note that in describing how attention is shifted, the theory of guided search explicitly (Wolfe, 2021, p. 1068) on the ideas of Koch and Ullmann (1985) about the WTA mechanism, which will be discussed in detail below.

The element selected by attention is placed in working memory, which also contains a guiding template and can determine the subsequent direction of attention. For example, when searching for a banana, attention is directed to target attributes using the templates "yellow" and "curved" (Wolfe, 2021, p. 1064).

To be identified as targets or rejected as distractors, objects selected by attention must be compared with target templates stored in the activated long-term memory (ALTM) fragment activated by the current task. The comparison helps to establish that the object is not just yellow and curved, but actually the banana that needs to be found. If there are only a few guiding templates in working memory, there can be many target templates; as an example, Wolfe cites the so-called hybrid search (Wolfe, 2012), see also (Angelhardt, Makarov & Gorbunova, 2021; Sapronov, Makarov & Gorbunova, 2023; Rubtsova & Gorbunova, 2022). These templates can be either specific (a ripe banana) or much more general (a fruit).

The binding and recognition of the object of attention is modeled as a diffusion process (Voronin, Zakharov, Tabueva & Merzon, 2020; Ratcliff, 1978), carried out at a speed of > 150 ms/element. Selection can occur more frequently if several elements are recognized simultaneously, albeit asynchronously; this makes controlled search a hybrid of sequential and parallel processes. For each target pattern stored in the ALTM, there is one diffuser (diffusion channel) that accumulates data (including noise) approaching the output threshold. When the data reaches the threshold, the search stops and either a true or false positive response is given. The search may also stop when the output signal accumulation threshold is reached, resulting in either a true or false positive response.

The accumulation threshold is adaptive, allowing feedback from previous presentations to program subsequent searches. Simulation shows that combining asynchronous diffusion with an output signal can reproduce the basic patterns of response times and errors obtained in a series of visual search experiments.

Thus, the theory of guided search explicitly describes the algorithm of attentive selection, closely resembling the theory of saliency. Thanks to this, it successfully overcomes the limitations of the theory of feature integration. In addition, it significantly expands the latter in terms of describing the algorithms of decision-making by the observer. The theory of guided search is developed mainly within the framework of the theoretical- informational approach and the traditional experimental-psychological paradigm of cognitive research. Saliency theory is at the intersection of cognitive and technical sciences and mainly describes the early stages of visual processing associated with the deployment of attention; modeling is an important part of it.

The theoretical foundations of mathematical and computer modeling of saliency were laid more than 40 years ago by the work of K. Koch and S. Ullman, which examines spatial shifts in attention and their possible neural mechanisms (Koch & Ullman, 1985). It should be noted that the term "saliency" had been used in psychology before, but as a more general concept that did not reflect the specifics of the work of a particular sensory system. Thus, as early as 1977, A. Tversky published a significant theoretical work formalizing the concept of "similarity" (Tversky, 1977) in set-theoretic terms. To summarize its content briefly, we can say that each object is characterized by a set of features, some of which are common to other objects, and some of which are distinctive and unique. Saliency (rather in the sense of "noticeability, significance") in Tversky is a property of a feature; it depends both on its physical characteristics stick (brightness, etc.), as well as from so-called diagnostic factors–contextual relevance and the importance of this feature for solving a specific task. Saliency occupies an important place in Tversky's theoretical constructs: thus, a more salient object is more likely to become a reference point in human judgments about similarity. The degree of similarity between objects **a** and **b** can be assessed on a scale **S** as:

$$S(a,b) = \phi f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

$$\phi, \alpha, \beta \geq 0,$$

where **A** and **B** are the sets of properties of **a** and **b**, respectively, and **f** is the saliency measure, which, like the parameters ϕ, α, and β, depends on the context and the task at hand. Thus, the saliency of an object can be determined within the framework of assessing the similarity of objects. A fairly simple interpretation of Tversky's ideas is given by B. Jules in his work (Julesz, 1986): saliency can be defined as a function (e.g., a ratio) of the number of unique and common features, or as a function of the number of unique features relative to their total number.

The concept of visual saliency itself was introduced by Koch and Ullman (Koch & Ullman, 1985) as a designation for the fundamental link in the organization of visual attention, combining information from individual feature maps into a general map containing measures of "conspicuity." The work was theoretical in nature and was largely based on the ideas expressed by Treisman and Gelade (1980), expanding on them in terms of explaining the algorithm for switching focal attention. Let us consider this article in more detail, as it has had a decisive influence on the entire field of attention research, while remaining virtually unknown in Russia.

The authors begin their article with arguments in favor of a two-level theory of human visual perception, which assumes the existence of a preattentive level, at which simple features are processed quickly and in parallel across the entire field of vision, and an attentional level. At the second level, the specialized focus of processing, i.e., the focus of attention, is directed at a specific location in the field of view, with the analysis of complex shapes and object recognition are associated with this level. If specific algorithms that solve problems such as shape analysis or object recognition in a specific location were performed in parallel, this would lead to a combinatorial explosion in the volume of required computations and a shortage of the necessary resources. The authors refer in particular to the criticism of the capabilities of perceptrons presented by M. Minsky and S. Papert in their well-known book (Minsky & Papert, 1971), which is of particular historical interest. Indeed, parallel processing in modern convolutional networks could hardly serve as a metaphor for the limited capabilities of the parallel stage of information processing in humans; however, the shallow fully connected perceptrons of those years were quite suitable for this role. In the end, the authors conclude that after a certain (parallel) preprocessing stage, the analysis of visual information continues in a sequence of operations, each of which is applied to a selected location or locations.

In presenting experimental evidence of selective attention, Koch and Ullman rely on both "psychophysical" (sic!) and physiological data. The existence of a moving specialized processing focus associated with foveal projections, but not identical to them, is confirmed by two classes of psychophysical experiments. First, there are the studies by Traiman and colleagues, in which "the search for a target specified by a single feature . . . , turns out to be parallel . . . , while the search for a conjunctive target defined in terms of several features . . . requires sequential, arbitrarily interrupted scanning among the presented distractors" (Koch & Ullman, 1985, p. 219). A number of studies devoted to the identification of visually detectable features also belong to this class of evidence. Thus, in their studies of texture discrimination, Julesz et al. showed that only a limited set of texton features can be detected in parallel (Bergen & Julesz, 1983–0029; Julesz, 1984). Secondly, there are a number of early studies using the spatial cueing paradigm (Bashinski & Bacharach, 1980; Eriksen & Hoffman, 1972; Posner, 1980; Remington & Pierce, 1984). Currently, there are several established names for tasks of this type: Posner cueing task, spatial cueing, Posner paradigm, cueing method, etc. (Gusev & Utochkin, 2012; Shevel & Falikman, 2022). Physiological data also support selective processing of visual information. Presenting a series of studies recording cellular activity, the authors conclude that "individual cells in certain parts of the visual system respond differently to identical physical stimuli, increasing their response as a function of the visual task being solved" (Koch & Ullman, 1985, p. 220).

As a result of their analysis, the authors formulate a number of fundamental questions about the mechanisms of selective processing. They are interested in what operations can be applied to selected locations, how this selection is carried out, and, in particular, how the change of locations is carried out.

Moving on to theoretical constructs, the authors first introduce the concept of early representation–a set of topographic cortical maps that encode visual information at the level of various elementary features, such as boundary orientation, color, disparity, and direction of motion. Each location in such maps has multiple feature dimensions. Probably, in accordance with evidence of the existence of spatial-frequency channels in the visual system (e.g., Campbell & Robson, 1968; Wilson & Bergen, 1979), there may be sets of maps with different resolutions for each individual feature. The maps contain neighborhood relations and local inhibitory connections (lateral inhibition), thanks to which locations that differ significantly from their surroundings can be detected at this early stage of analysis. Thus, the maps "signal" the conspicuity of a section of the visual scene.

We are talking specifically about conspicuity, not saliency. Saliency arises at the next stage of processing as a separate perceptual mechanism. This explains the need to directly transfer the term "saliency" into Russian; attempting to translate it could lead to confusion when naming the levels of processing.

When attention is focused on a particular location, the features present in it must be transferred to a higher, more abstract and non-topographic level of representation. The authors note that this formulation of the question does not contradict the idea of hierarchical information processing in the cortex; we also note that it is consistent with the basic tenets of feature integration theory. How is the location for attention selected? How is high-dimensional feature information represented in early representation processed?

The authors suggest that the saliency of a location in the visual scene determines the level of activity of the corresponding elements in various feature maps, with different maps encoding saliency within a specific feature dimension. All this diverse information is combined thanks to a saliency map, which is a single global measure of saliency that, like feature maps, has a topographical structure. The authors do not describe the exact nature of the process of combining feature maps, assuming that it, still being part of the early visual system, "encodes the saliency of objects in terms of simple properties such as color, direction of motion, depth, and orientation" (Koch & Ullman, 1985, p. 221). It was this uncertainty that served as a starting point for a whole new direction of research in the future. Note that the authors also allowed for the possibility of modulating influences on the saliency map from higher cortical centers; in the future, such influences would begin to be implemented in models of top-down saliency.

The central place in Koch and Ullmann's theoretical constructs is occupied by the main link of attentional selection, which was explicitly absent in the theory of feature integration–the WTA ("winner takes all") network (Feldman, 1982), which is responsible for selecting the location for focal attention, the properties of which are then transferred to the "central representation"; it works with a saliency map.

The WTA mechanism can be viewed as equivalent to a maximum search operator operating on the elements of the saliency map xi ; in a neural network, xi can be interpreted as the electrical activity of an element at location i. WTA maps a set of input elements to an equivalent set of outputs yi according to the following rule:

$$y_i = 0 \; if x_i < \max_j x_i$$
$$y_i = f(x_i) \; if x_i = \max_j x_j,$$

where f is any increasing function of x_i or a constant. Thus, all output elements except one, corresponding to the most active input element, are set to 0.

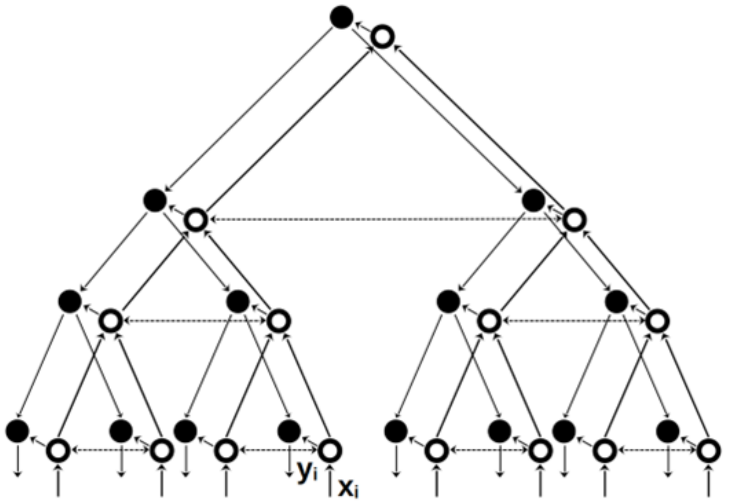
If we disregard the "hardware" features of the brain substrate of computations, building a WTA network seems to be a fairly simple task. The authors consider a number of possible implementations of the network, both fully sequential, which is unacceptable due to its extremely slow operation, and highly parallel, characterized by too many connections between processing elements and the inability to process an arbitrary number of inputs. Based on this, the authors formulate two biologically plausible assumptions, building on them possible implementations of WTA:

1. "With the exception of some distant excitatory connections, most of them, both excitatory and inhibitory, are local" (Koch & Ullman, 1985, p. 222).
2. "Each elementary processing element performs only simple, well-defined operations, such as addition or multiplication. In particular, basic processing elements are incapable of processing any symbolic information, such as addresses".

There are two such implementations in total, and the authors clearly prefer the second one. This WTA implementation has a hierarchical pyramidal structure and operates in a highly parallel mode. First, the maxima among m elements from the input set of size n are calculated. At the next level of the hierarchy, the process is repeated for n/m input elements; This continues until the pyramid of comparisons closes on the last element, which displays the global maximum. However, both the absolute value of the maximum and its location are important for the selection process. It is determined using a second pyramid of additional elements, in which information is distributed in reverse order. Each additional element is associated with an element of the main pyramid and is activated only when it receives simultaneous excitation from its main element and from an additional element located at a higher level. "Since at each level the most activated element of the main pyramid in suppresses the activity of the other $m-1$ main elements in a local comparison, associated additional elements, as well as all additional elements in the lower branches, will never be activated" (Koch & Ullman, 1985, p. 223). Fig. 2 shows a possible example of a WTA network implementation with $n=8$inputs and $m=2$ comparator elements. The number of ascending and descending time´ x computational steps for such a network should not exceed $2\log_m n$, the network contains no more than $2nm/(m-1)$ elements. It is assumed assumes that the input values do not have to be exactly the same.

Figure 2

Second implementation of the WTA network with $n = 8$ input elements. Local comparison is performed between m= 2 elements. The main elements are shown in light color, the additional ones in black; x(i)corresponds to the maximum at the network input, y(i)corresponds to the network response to the detected maximum. According to (Koch & Ullman, 1985).



Local comparison is performed between$m = 2$ elements. Primary elements are shown in light, secondary elements in black; $x_i$corresponds to the maximum at the network input, ,y_i the network response to the detected maximum. According to (Koch & Ullman, 1985).

The authors provide estimates according to which only a small portion of the available visual neurons is sufficient for the implementation of the WTA network in a living system (primates, cats). Presumably, large-cell systems, such as the Y-path in cats, are well suited for the role of the WTA substrate.

How does the change in locations captured by attention occur across the visual field? Two mechanisms are possible here, local and central, acting through modification of the saliency map. The local mechanism can be implemented through adaptation and weakening of the active location in the saliency map over time; the most active element is locally inhibited, for example, after a certain time interval. The central mechanism activates an inhibitory signal from the central representation, where the information was previously received. There is no contradiction between the existence of these mechanisms, and they can operate simultaneously; it is likely that the local mechanism is constantly engaged, while the central mechanism is activated when there is an impulse to shift attention arbitrarily (Posner, 1980). ´Both of these mechanisms implement long-term inhibition of the selected element of the saliency map, preventing a repeat visit to the corresponding location for a certain period of time–the so-called inhibition of attentional return return (Utochkin & Falikman, 2006; Posner, Cohen, & Rafal, 1982).

The attentional selection mechanisms proposed by Koch and Ullmann, based on saliency maps and WTA, enable them to offer their interpretation of the effects of parallel and sequential search, as well as the camouflage of a specific object by others (Treisman, 1982). If the target has a salient feature that distinguishes it from its neighbors, WTA will immediately determine its location, and the target will be detected in a time that does not depend on the number of distractors. If the target is determined by a combination of features, the saliency map will have many local peaks, "in the worst case, as many as there are objects presented" (Koch & Ullman, 1985, p . 224). If no additional optimization strategy is applied, WTA will go through them; thus, to successfully complete the search, it will be necessary to view an average of $n/2$ of the presented objects. Thus, an object "pops out" because, due to its saliency, it is the first one to be visited, and parallel and sequential searches are not fundamentally different processes. As for masking, there are two different strategies: you can reduce the visibility of an object by blending it with its surroundings (this is roughly how military camouflage works), or you can place it among very visible objects. In both cases, the activity of the saliency map at the point corresponding to the target object will decrease relative to its surroundings.

What is the additional optimization strategy that allows, in a significant number of cases, to avoid the need for a complete search of objects in the visual scene? The authors believe that such a strategy can be based on the rules of proximity and similarity priorities, roughly corresponding to the phenomena of perceptual grouping and the Gestalt principles of the same name. Thus, searching for a target around a selected location will be more successful if the selection mechanism's preferences are shifted toward neighboring locations. As experimental confirmation of the priority of proximity, the authors cite studies demonstrating the dependence of the probability of target detection on proximity to the location on which attention is focused (F. L. Engel, 1971, 1974). The search for objects with a common distinguishing feature will improve if locations with properties similar to those represented in the current location become preferred. This is partially confirmed by the results that were in press at the time of writing (Geiger & Lettvin, 1986): the demonstration of a figure at the fixation point makes the same figure appearing elsewhere in the field of view in the same presentation salient.

The simplest way to implement proximity priority within the WTA mechanism is to enhance all elements in the saliency map that are adjacent to the currently selected one. "The output of the WTA mechanism associated with the selected location increases the saliency of nearby elements in the saliency map by an amount depending on the distance between that location and its surroundings, thereby facilitating a shift in the focus of processing to nearby locations," which "is equivalent to the assertion that there is attractive potential around each selected location" (Koch & Ullman, 1985, p. 224).

The priority of similarity can be implemented as follows. When triggered, the WTA mechanism initiates interactions within individual maps of signs at the level of early representation, thanks to which maps containing currently selected features become more visible in the vicinity of the selected location. This process does not involve interaction between feature maps or their precise topographical reference to each other. If an object with a red horizontal line is selected, the neighboring locations in the "red" and "horizontal" feature maps will be enhanced; the focus of attention is more likely to shift to them.´ The process that ensures the priority of similarity acts in opposition to the initial priority of salient locations, which arises due to lateral inhibition within feature maps; various options for the interaction of these processes are possible.

These are, in general terms, the main theoretical positions put forward by Koch and Ullmann in 1985. The first computational models of saliency appeared much later, in the mid-1990s (Baluja & Pomerleau, 1994; Itti, Koch, & Niebur, 1998; Milanese, 1993; Tsotsos et al., 1995); as they improved, they began to gain practical significance. Let us now consider the main results obtained within the framework of various approaches to modeling.

Discussion

Computational saliency Models

Approaches to saliency modeling can be broadly divided into traditional and neural network approaches. Thanks to the use of modern neural network architectures, primarily convolutional ones, all records for model training quality have been broken in recent years (Borji, 2019). The success of neural network models is facilitated not least by the increase in the volume of publicly available data from eye-tracking studies and the emergence of standardized and relatively easy-to-use neural network modeling tools. Let us consider these approaches in more detail, starting with the traditional ones that have had the greatest impact on the subsequent development of the field.

The model developed by Laurent Itti, Christoph Koch, and Ernst Niebur served as the basis for many subsequent models; it also serves as a benchmark for comparing them (Borji & Itti, 2013). The model analyzes intensity, color, and orientation. In the first stage, the input color (r, g, b) image 640x480 in each of the corresponding channels is represented as a Gaussian pyramid (9 scales from 1 :1 to 1 :256 with an octave step). The intensity representation of the image $I = (r + g + b)/3$ is used to create the pyramid $I(\sigma)$, where $\sigma \in [0..8]$ is the scale. It is also used to normalize the primary color channels r, g , and b , which is used to separate color hue from intensity. Since hue changes are not perceived at low brightness, normalization is applied only where I is greater than $1/10$ of its maximum across the entire image; in other locations, pixel values are set to zero.

Local feature maps are calculated using a set of linear central-peripheral operators, which are implemented in the model as a point-by-point difference between fine high-frequency and coarse low-frequency scale representations (denoted by $\ominus$): the center is represented by pixels at scale $c \in \{2,3,4\}$, and the neighborhood is the corresponding pixels at scale$s = c + d$ where $d \in \{3,4\}$. Six intensity maps are calculated as

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|.$$

Based on the primary normalized color channels, four new broadband channels are created:

- red: $R = r - (g + b)/2$
- green: $G = g - (r + b)/2$
- blue: $B = b - (r + g)/2$
- yellow: $Y = (r + g)/2 - |r - g|/2 - b$.

Negative values are set to zero. Pyramids are created from these channels $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ and $Y(\sigma)$.

Sets of maps for color channels are created similarly to intensity maps, while channels with double color opposition are modeled (Hohlova, 2012; S. Engel et al., 1997): the centers of the receptive fields of neurons are excited by one color (e.g., red) and inhibited by another, while the opposite occurs at the periphery. Maps modeling dual color opposition in the primary visual cortex of humans (green/red ($\mathcal{RG}$) and blue/yellow ($\mathcal{BY}$)), are calculated using the formulas

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|,$$

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|.$$

Local orientation information is extracted from I using an oriented Gabor pyramid $O(\sigma,\theta)$, where $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Orientation feature maps $\mathcal{O}(c,s,\theta)$ encode local differences in orientation between the center and the periphery, represented by different scales:

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|.$$

Thus, a total of 42 feature maps are created: 6 for intensity, 12 for color, and 24 for orientation.

Combining feature maps into conspicuity and saliency maps is problematic: different modalities have different dynamic ranges and use different feature extraction mechanisms, making them difficult to compare. In addition, salient objects represented on only a few feature maps may be masked by noise or less salient objects represented on a larger number of maps. In the absence of a mechanism in the model that provides top-down control, the authors propose using the map normalization operator $\mathcal{N}(.)$, which would increase the global role of those that contain a small number of strong activity peaks and would lower it for those that contain a large number of peaks of comparable strength. The application of $\mathcal{N}(.)$, involves:

1. bringing map values to a single fixed range $[0..M]$, to eliminate modality-specific amplitude differences;
2. searching for the global maximum of the map M and calculating the average $\overline{m}$ of all its local maxima;
3. global multiplication of the map by $(M - \overline{m})^2$.

The authors use the model of cortical mechanisms of lateral inhibition to explain how the operator works (Cannon & Fullenkamp, 1996): when $M - \overline{m}$ is sufficient, the most active location stands out sharply, and the map becomes more important; if the difference is small, the map contains nothing unique and turns out to be insignificant. Feature maps are combined into three saliency maps $\overline{I}$, $\overline{C}$ and $\overline{O}$, for intensity, color, and orientation, respectively. Saliency maps are created by summing all the maps of the Gaussian pyramid after bringing them to a single scale with $\sigma = 4$; this operation is referred to by the authors as $\bigoplus$:

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}(\mathcal{I}(c,s)),$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))],$$

$$\overline{O} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}(\mathcal{O}(c,s,\theta))).$$

The process of calculating $\overline{O}$ involves creating four intermediate maps by combining six feature maps for each θ , and then combining them into a single saliency map.

The authors explain the creation of three independent channels $\overline{I}$, $\overline{C}$ and $\overline{O}$, and their separate normalization by the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. The three saliency maps are normalized and summed into the final input $\mathcal{S}$ of the $\mathcal{SM}$ saliency map:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\overline{I}) + \mathcal{N}(\overline{C}) + \mathcal{N}(\overline{O})).$$

At each moment in time, the maximum activation of the SM map determines the most salient location in the image on which attention should be focused. To determine the point to which the model should switch next, one could simply select the most active location on the map. However, based on considerations of biological plausibility, the authors model the saliency map as a two-dimensional layer of leaky integrate-and-fire neurons on the $\sigma = 4$. The model of such neurons includes a single "capacitor" that accumulates charge from the synaptic input, leakage conductance, and threshold voltage. When the threshold is reached, an "action potential" (prototypical spike) is generated, and the charge of the "capacitor" is reset to zero. The maximum activation of the map enters a biologically plausible two-dimensional WTA neural network, in which synaptic interactions between elements ensure that only the most active location remains, while all others are suppressed (here the authors refer us, among other things, to the previously discussed work (Koch & Ullman, 1985).

Neurons in the SM receive excitatory input from $\mathcal{S}$ and are independent of each other; therefore, their potential in more salient locations increases faster (these neurons are used as pure integrators and do not fire continuously). Each SM neuron excites its corresponding WTA neuron. All WTA neurons also change their state independently of each other until one ("winner") is the first to reach the threshold and fire. This triggers three simultaneous mechanisms:

1. the focus of attention shifts to the location of the winning neuron;
2. global inhibition is triggered and completely suppresses (resets) all WTA neurons;
3. In SM, in the area corresponding to the position and size of the new focus of attention, local inhibition is temporarily activated; this not only leads to dynamic shifts in focus, allowing the next most salient location to subsequently become the winner, but also prevents the focus of attention from immediately returning to the previously visited location.

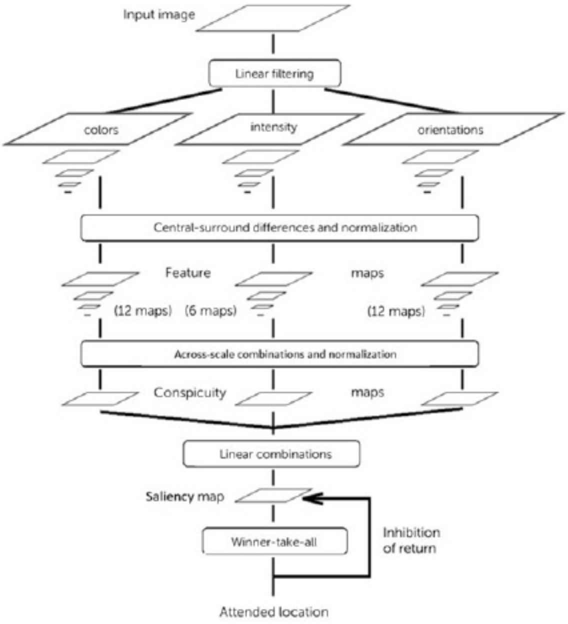
Such "inhibition of attention return" has been described in studies of human vision (see, e.g., (Utochkin & Falikman, 2006)). In addition, Koch and Ullman's "proximity preference" rule is modeled (Koch & Ullman, 1985): to slightly reorient the model toward finding the next salient location close to the previously visited one, in the SM, in the area corresponding to the position and size of the new focus of attention, local inhibition is temporarily activated; this not only leads to dynamic shifts in focus, allowing the next most salient location to subsequently become the winner, but also prevents the focus of attention from immediately returning to the previously visited location: in order to slightly reorient the model to search for the next salient location close to the previously visited one, in SM, in the immediate vicinity of the current focus of attention, a small excitation is temporarily activated.

Since this saliency model does not take into account top-down "top-down" controle, the focus of attention is a simple disk, the radius which is constant and equal to $\frac{1}{6}\min(h,w)$, where h, w are the height and width of the input image, respectively.

´The time constants, conductivity values, and thresholds of the simulated neurons were chosen so that the focus shifted from one salient location to another in approximately 30–70 ms, and the previously visited location was suppressed for approximately 500–900 ms, which corresponds to psychophysical data (Posner & Cohen, 1984). The difference in the relative magnitude of these delays was sufficient to ensure complete scanning of the image and prevent looping on a limited number of locations. ´All tuning parameters are fixed in the author's implementation of the model in C++, and with them, the system demonstrates temporal stability on all test images. A generalized diagram of the model is shown in Fig. 3.

Figure 3

General diagram of the saliency model by L. Itti, K. Koch, and E. Niebur. Adapted from (Itti et al., 1998).



The review by Ali Borji and Laurent Itti (Borji &amp; Itti, 2013), which essentially summarizes the development of saliency modeling up to the moment of widespread interest in deep learning technologies, covers more than fifty models published between 1998 and early 2012. The authors analyze 52 saliency models that primarily consider ascending attention, although this analysis does not include developments known to them (Baluja & Pomerleau, 1994; Milanese, 1993; Tsotsos et al., 1995) presented before 1998, i.e., before the publication of "the first complete implementation and verification of the Koch and Ullmann model proposed by Itti et al." (Borji & Itti, 2013, p. 186). The review also analyzes works presenting more generalized models of attention with top-down control—there are 11 of them, two of which were proposed before 1998 (McCallum, 1996; Rao, Zelinsky, Hayhoe, & Ballard, 2002). It probably makes no sense to list all the models considered here; however, the theoretical generalizations made by the authors in the course of their analysis, a summary of which is presented below, are particularly interesting. The authors highlight the following properties of the models that are important for categorizing and understanding their features:

1. bottom-up and top-down control. Models can represent predominantly ascending attention control factors based on certain characteristics of the visual scene, or

descending factors (knowledge, expectations, reinforcement, current goals, etc.), or take both into account. At the same time, they differ in:

– the features used. Both individual low-level features (color, orientation, etc.) and fairly complex object properties can be taken into account. In cases where the model includes top-down control, a mechanism for adjusting feature detectors can be used. Models that process features are closely related to purely computational methods of object detection; cognitive modeling and computer vision enrich each other;

– the degree of scene context consideration. It is known that with very short exposures (80 ms or less), the observer is able to grasp the main content ("gist") of the scene. Its representation does not contain a large number of details of the objects presented in it, but it can provide sufficient information for coarse discrimination (e.g., inside or outside a room). The influence of context is also evident in the speed of object detection and in the characteristics of eye movements. Traditional computational models that take into account the main content of a scene typically use filtering (including biologically based methods such as central-peripheral filtering and Gabor filters) or spectral methods to extract features, the dimensionality of which is then reduced using principal component analysis (PCA) independent component analysis (ICA), or cluster analysis. The result is a vector of values ("gist vector") that characterizes the scene. The authors of the review note that at the time of writing, the popularity of this approach in computer vision was growing.

– taking into account the requirements of the task. The task greatly influences the distribution of attention, and scenes can be interpreted based on the needs that arise to meet the task requirements. When solving complex tasks, there is a strong connection between visual cognition and eye movements. Thus, during visual control, most fixations are directed at areas relevant to the task. Eye movements often reveal the solution algorithm used by the subject. In particular, in the block copying task (Ballard's paradigm, for more details, see (Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Pook, & Rao, 1997; Hayhoe & Ballard, 2005)), which involves the test subject reproducing a structure from elementary "building" blocks of different colors, the test subjects first selected the target block in the original structure, confirming its position, and then fixed their gaze on the workspace to place the corresponding block in the correct place. The authors also provide a list of studies in which activities in natural conditions were investigated in a similar manner.

The authors of the review note that ascending and descending attention combine to control our attention, providing several options for implementing the rules for integrating these processes.

2. only space or space and time. Models can take into account the movement of objects and predict attention shifts between objects in a static or dynamic scene;

3. overt and covert attention. Models can describe both overt and covert attention, but the degree to which they account for covert attention is difficult to assess due to the complexity of measuring it;
4. objects or spatial locations. Given that there are grounds for distinguishing between feature-based attention and object-based attention, models may give preference to one of these types;
5. features used in the model. Many models use traditional features used in integration theory; however, there are many others, such as mathematically constructed (wavelets, PCA, ICA), geometric, etc.;
6. stimuli and task type. Since real empirical data is needed to test the model, the authors identify two grounds for distinguishing models based on the stimuli used in data collection: static/dynamic and artificial/natural. The type of task solved by the observer is also important. It can be free viewing, visual search, or an interactive task;
7. metrics used for evaluation. When evaluating a model, its prediction is usually compared with an empirically obtained result (ground truth); often, various versions of gaze fixation maps are used as such a result. Depending on the map and the type of result produced by the model (fixation points, two-dimensional probability distribution, etc.), several modifications of the area under the curve, normalized saliency of the gaze path, Kulback-Leibler metric, Pearson's correlation coefficient, etc. can be used. A detailed discussion of various metrics can be found in a more recent work (Bylinskii, Judd, Oliva, Torralba, & Durand, 2017);
8. eye movement datasets used. At the time of the publication of the review by Itti and Borji, eye movement data recorded while viewing static images (Bruce & Tsotsos, 2005; Judd, Ehinger, Durand, & Torralba, 2009) and videos (Marat et al., 2009) were freely available. Many authors used their own data to train and test models, which eventually became available to other researchers;
9. Models can be classified based on how saliency is calculated. For example, a model can be based on neuron-like calculations, or it can use formal high-level approaches. The authors note that some models fall into several categories at once, but nevertheless use a simple single-level classification in the future:

– cognitive models. Almost all models of attention were created under the influence of cognitive concepts. However, the authors include in this class those models that are more closely related to psychology or neurophysiology; the author of this review believes that this may be a matter of substantive connection, since the algorithms used in these models intersect in one way or another with psychological and/or neurophysiological concepts;

– Bayesian models. "In these models, prior knowledge (e.g., the context of the scene or its gist) and sensory information (e.g., target features) are probabilistically combined according to Bayes' rule (e.g., to detect an object of interest)" (Borji & Itti, 2013, p. 194). These models are capable of learning from data and generalizing various factors;

– models based on decision-making theory. These models are based on the idea that visual attention should be managed in an optimal way in the context of the current task; they can be based on very different algorithms (both biologically based and purely computational);

– models based on information theory. These models are based on the assumption that salient areas are the most informative in terms of the amount of information they contain. Computationally, these models are based on comparing various statistical estimates of image regions (entropy, distribution parameters, etc.);

– Graphical probability models. "Graphical models can be viewed as a generalized version of Bayesian models" (Borji & Itti, 2013, p. 197). Such models use graphs that represent the structure of conditional independence of random variables; eye movements are viewed as a time series. Due to the existence of hidden variables that influence the formation of eye movements, solutions such as hidden Markov models (HMM), dynamic Bayesian networks (DBN), and conditional random fields (CRF);

– models based on spectral analysis. This group of models is based on the analysis of image properties, often with scaling, represented in the frequency domain (amplitude and phase spectrum);

– models based on pattern classification. These models use machine learning methods such as support vector machines (SVM), regression, etc. Training is carried out on specially labeled data (for example, divided into areas, each of which is marked as salient or non-salient);

– Other models. A fairly extensive and highly blurred "class" of models characterized by originality and based on a wide variety of computational solutions.

Based on these properties, the authors of the review have compiled an extremely useful summary table of the models they have considered (Borji & Itti, 2013, p. 201), allowing the reader to quickly navigate the vast array of rather complex developments and find the necessary bibliographic information. Each of the listed properties is represented by a column in the table, with the models known to the authors listed in rows; the cells contain symbols that indicate whether a model has a particular property. Thus, using the table, one can quickly determine that the classical model of Itti, Koch, and Niebur (Itti et al., 1998) discussed earlier is ascending, spatial, rather than spatiotemporal, static; dealing with natural stimuli and the task of free viewing, based on spatial locations rather than objects, taking into account only simple features (color, brightness, orientation), cognitive; data for training the model were not used.

Neural network models of saliency

Moving on to the description of saliency models based on deep learning methods, we cannot fail to mention the existence of a remarkable review published by A. Borji in 2021 (Borji, 2021), but available as a preprint since 2019 (Borji, 2019). I would like to recommend this document to interested readers as a valuable source of reference information on neural network models and datasets created over the past decade, on the metrics used, and on methods for evaluating model performance. Given the existence of this high-quality review, the author of this text (D. Ya.) sets himself two fairly modest tasks: to acquaint the reader with the history and logic of the development of the field using the example of the work of one of the most successful research groups working in the field of saliency modeling; to examine the models created after the publication of Borji's review and attempt to identify and summarize their characteristic features.

The work of A. Krizhevsky, I. Sutskever, and J. E. Hinton (Krizhevsky, Sutskever, & Hinton, 2012) sparked another revolution in artificial intelligence research, reviving widespread interest in deep learning neural networks, which had faded somewhat due to the rapid development of machine learning approaches such as kernel methods and decision trees at the turn of the century (see, e.g., (Chollet, 2023)). The model, later named AlexNET, won a decisive victory at the annual ImageNet competition in 2012, achieving a record performance of 83% in the classification of 1,000 object categories. The use of the then-novel multilayer convolutional architecture and graphics processing units allowed researchers to achieve impressive results in the following years, including in the modeling of visual saliency.

As early as 2014, a group of researchers from the University of Tübingen (Bethge Lab) developed the DeepGaze I model (Matthias Kümmerer, Theis, & Bethge, 2015), which used weights from the neural network of A. Krizhevsky et al. 2015). The use of transfer learning technology allowed the authors to achieve a significant increase in performance compared to previously created models. Thus, the correlation between predictions and tracking data on the MIT300 dataset is 0.6144. The model used the outputs of the convolutional layers of AlexNET, which were linearly combined with different weights. The resulting layer was filtered (convolution with a Gaussian kernel), then a weight matrix implementing a center bias correction was added to it elementwise. In this form, the result was fed to the softmax layer, at the output of which the distribution of fixation probabilities was formed. To stimulate sparsity, l1 regularization of weights was applied in the model.

In 2017, a new version of the model appeared, DeepGaze II (M. Kümmerer, Wallis, Gatys, & Bethge, 2017). It used the convolutional part of VGG-19 (Simonyan & Zisserman, 2015) as its base; information was extracted from the conv5_1, relu5_1, relu5_2, conv5_3, and relu5_4 layers. The trainable part was made more complex (4 convolutional layers 1x1), but otherwise the model was similar to the previous one. The model demonstrated very high performance at the time: the correlation between the empirical MIT300 data and the forecast was 0.7703.

In parallel with it, the DeepGaze ICF model was created, in which, instead of the basic part in the form of network layers that were pre-trained to recognize objects, operations for extracting exclusively low-level features were used. Calculations were performed for brightness and two color difference components in five scales (Gaussian pyramid) for brightness and contrast, respectively; thus, 30 low-level feature maps were generated at the output. This model achieved better performance (correlation of 0.5876 on MIT300) than all models that did not use features from neural networks pre-trained to recognize objects, which, according to the authors, makes it a reliable basis for assessing the usefulness of high-level features. Thanks to this model, the authors found that some fixations are much better predicted by low-level features.

The DeepGaze IIE model (Linardos, Kümmerer, Press, & Bethge, 2021), introduced in 2021, is an improved version of DeepGaze II. The trainable part of the network has been made deeper, and ReLU activations have been replaced with norm and softplus. Training was performed on the Salicon and then MIT1003 datasets. The main change concerned the base network: the original VGG-19 could be replaced with other deep networks trained on the ImageNet dataset (ResNet50 (He, Zhang, Ren, &amp; Sun, 2015), EfficientNet85 (Tan & Le, 2020), etc.). According to MIT/Tübingen Saliency Benchmark), the highest correlation between the prediction and empirical fixation maps was 0.8242; in fact, this is the best model tested to date and presented on the website. However, the authors continue to create new versions of the model.

In 2022, DeepGaze III was introduced (Matthias Kümmerer, Bethge, & Wallis, 2022; Matthias Kümmerer, Wallis, & Bethge, 2022), which includes a spatial prediction module that takes into account the influence of scene content on fixation location, and a scan history module that identifies the influence of earlier fixations and, consequently, the dynamics of gaze trajectory. The first module broadly replicates previously developed spatial models; the second uses information about four or fewer previous fixations to predict the current fixation, which is represented as maps of three features: distance to the current fixation, as well as x and y displacements. Information about previous fixations made by the subject is processed in this module and then combined with the spatial map in the fixation selection network. The final prediction is blurred, combined with the central offset correction weights, and converted into a probability distribution using softmax. Judging by the AUC= 0.906 and NSS= 2.957 values reported by the authors, obtained on MIT300 (the correlation value is not given), the model demonstrates the highest performance of those previously presented, but data on it on the MIT/Tübingen

Saliency Benchmark is not yet available. The approach used by the authors allows us to investigate the influence on perceptual saliency not only of the physical properties of the image and the task, but also of previously produced fixations.

The idea of processing features extracted from layers of a convolutional neural network trained to recognize objects is also used by the authors of the TranSalNet model (Lou, Lin, Marshall, Saupe, & Liu, 2022). When developing the model, they set themselves not only the task of obtaining maximum results, but also sought to bring the architecture of the artificial network closer to the human perceptual system. First, the image is fed into a convolutional encoder. To obtain multi-scale representations, three sets of feature maps with different spatial dimensions are extracted from the encoder. Due to the inductive biases inherent in convolutional architectures, the extracted image representations do not contain contextual information at a large scale, which potentially makes the saliency model less human-like. The authors draw the reader's attention to the fact that the human visual system is capable of capturing both local and global information. The authors stress that the saliency model is not as human-like as it could be; they emphasize that the human visual system is capable of capturing both local and global information. Therefore, to obtain a prediction that is more relevant from the point of view of perception, these feature maps are passed through three encoder transformers (Vaswani et al., 2023), which allows us to obtain global feature maps with improved context information transfer. The encoder transformers contain a multi-head self-attention layer and a multilayer perceptron. Then, a convolutional decoder combines the feature maps to construct a saliency prediction. The model demonstrates performance comparable to DeepGaze: when using DenseNet-161 (Huang, Liu, Maaten, &amp; Weinberger, 2018) as the base network, the correlation between the prediction and the MIT300 data is 0.8070; with ResNet-50, the correlation decreases slightly (0.7991).

Despite their significant capabilities for forming representations of image elements, feedforward convolutional neural networks can ignore their internal connections and lack the potential advantages provided by the use of feedback in visual tasks. This also applies to saliency modeling. Given this circumstance, the authors of the SalFBNet model (Ding, İmamoğlu, Caglayan, Murakawa, & Nakamura, 2022) propose a convolutional architecture with feedback and recursion. The proposed model can form multiple contextual representations using a recursive path from higher-level feature blocks to lower-level layers. To address the problem of training data scarcity, the authors use a special approach to knowledge transfer, creating a large-scale training set using pre-trained saliency models listed on the MIT/Tübingen Saliency Benchmark website. First, they train the proposed model on the artificial data obtained in this way, then retrain it on real gaze fixations. In addition, to facilitate training their feedback model, the authors propose a new loss function, which they call sFNE (selective fixation and nonfixation error). Numerous experimental results show that SalFBNet with fewer parameters achieves competitive results in publicly available saliency model tests, which indicates the effectiveness of both the feedback model itself and the use of artificial data for pre-training. SalFBNet ranks second in performance after DeepGaze IIE (correlation with MIT300 data 0.8141).

The Saliency TRansformer (SalTR) model (Dahou Djilali, McGuinness, & O'Connor, 2024) is based on a new approach to predicting saliency in images, using parallel decoding in transformer networks to train the network exclusively on fixation maps. To overcome the optimization challenges for discrete maps, models are typically trained on continuous maps. The developers of SalTR attempt to build an experimental computing system that generates saliency datasets. The authors' approach treats saliency estimation as a direct prediction problem using a global loss function that predicts individual fixations through bilateral matching and a transformer-encoder-decoder architecture, with a ResNet50 base network at the input. Using a fixed set of learned fixation queries, cross-attention processes image feature information to directly infer fixation points, which distinguishes this development from other modern models. The authors note that their approach achieves estimates comparable to other modern approaches in the Salicon and MIT300 tests. Thus, the implementation of SalTR-Small provides correlations between predictions and original samples at the level of 0 .84 and 0 .7 for Salicon and MIT300, respectively, while SalTR- Base provides correlations of 0 .87 and 0.75. The use of deformable convolutions in the models increases the similarity to 0.86 and 0.76 (small) and 0.89 and 0.8 (base), respectively. Thus, SalTR is indeed one of the best modern models of visual saliency.

Modeling of visual saliency is also developing in the direction on video stream processing. In their work (Droste, Jiao, & Noble, 2020), the authors point out that saliency modeling for images and videos is considered in the current literature on computer vision as two independent tasks. And while modeling for images is a well-developed problem, and progress in this area is slowing down, as seen in the SALICON and MIT300 benchmarks, saliency models for video have recently shown rapid growth in the DHF1K benchmark (Wang et al., 2021). The authors ask whether it is possible to approach saliency modeling for images and videos using a single model with mutual benefits. In their opinion, the key prospects for joint modeling are provided by the application of domain shift (adaptation of an AI system to use in a new area and/or applying to new data) both between saliency data for images and for videos, and between different sets of video data. In addition to an improved algorithm for creating trained Gaussian priors (correction for gaze shift to the center), four new domain adaptation methods are proposed to solve this problem: domain-adaptive prior values, domain-adaptive fusion, domain-adaptive smoothing, and recurrent network bypass. These methods are integrated into a "simple and lightweight"

(Droste et al., 2020, p. 1) UNISAL network with an "encoder-recurrent block-decoder" architecture, trained on saliency data for both images and videos. The training results are evaluated on the DHF1K, Hollywood-2, and UCF-Sports video datasets, as well as on the SALICON and MIT300 static datasets. With the same set of parameters, UNISAL achieves the highest performance at the time of publication on all saliency datasets for video and is on par with the best models in tests on image data (correlation with MIT300 data is 0.7851); Compared to all competing models using deep learning, the execution time is reduced by 5–20 times, and the model itself is smaller. The authors also conduct retrospective analysis and ablation studies (studies of the role of an AI system component by disabling it), which confirm the importance of domain shift in modeling.

Characteristics of Modern Deep Learning Saliency Models

1. Modular neural network architectures with replaceable modules.
2. Knowledge transfer: leveraging pre-trained networks and artificial datasets for pre-training.
3. Domain adaptation: extending models across domains, e.g., images and videos.
4. Beyond classical convolution: use of recurrent paths, self-attention, feedback loops, and transformers.
5. Modular manipulation for ablation studies, enabling analysis of each component’s contribution.

Conclusion

A considerable amount of time passed between the publication of Koch and Ullmann’s seminal article (1985) and the practical testing and implementation of their ideas. Early research focused primarily on the algorithm for forming the initial saliency map, while many details of its construction were only briefly mentioned in the original work. The first, traditional stage of saliency model development was characterized by a wide variety of computational methods and approaches. Some of these solutions were well-aligned with psychological and neurophysiological data. At this stage, visual saliency models were largely “transparent” in terms of internal structure, making them especially valuable for comparison with theoretical models from cognitive science. With the rise of machine learning methods–such as Bayesian classifiers and support vector machines–particularly in the first decade of the 21st century, some conventional models began to resemble “black boxes.” This trend intensified dramatically after the 2012 revolution in neural network technology, though it also brought impressive gains in performance. There is hope that, as tools for analyzing the specific algorithms learned by neural networks improve, the contents of these “black boxes” will become more interpretable. Optimism is also supported by the growing volume of publicly available data for training saliency models, as well as a clear understanding in the research community of the importance of task type (e.g., free viewing, visual search) and task characteristics when collecting such data.

As effective computational approaches have matured, the literature has increasingly explored practical applications of saliency models, including computer vision (Medioni & Mordohai, 2005), engineering psychology and usability studies (Sun et al., 2019), medical image analysis (Arun et al., 2020; Jampani et al., 2012), and video compression (Gitman et al., 2014; Lyudvichenko et al., 2017). The first commercial solutions are also emerging. Thus, the modeling of visual saliency has now acquired significant practical relevance, enabling both the simulation of attention for technical purposes and the prediction of attentional shifts in humans.

References

Angelgard, A.N., Makarov, I.M., & Gorbunova, E.S. (2021). ROLE OF THE Category Level in Hybrid Search. *Voprosy Psychologii,* 2, 148-158. https://www.elibrary.ru/item.asp?id=46548586

Velichkovskij, B. M. (2006). *Kognitivnaya nauka. Osnovy psixologii poznaniya.* Smysl.

Arun, N. T., Gaw, N., Singh, P., Chang, K., Hoebel, K. V., Patel, J., Gidwani, M., & Kalpathy-Cramer, J. (2020, May 29). *Assessing the validity of saliency maps for abnormality localization in medical imaging.* http://arxiv.org/abs/2006.00063

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience*, 7(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(4), 723–742. https://doi.org/10.1017/s0140525x97001611

Baluja, S., & Pomerleau, D. (1994). Using a Saliency Map for Active Spatial Selective Attention: Implementation & Initial Results. *Proc. Advances in Neural Information Processing Systems,* 451–458.

Bashinski, H. S., & Bacharach, V. R. (1980). Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations. *Perception & Psychophysics*, 28(3), 241–248. https://doi.org/10.3758/bf03204380

Bergen, J. R., & Julesz, B. (1983–29C.E.). Parallel versus serial processing in rapid pattern discrimination. *Nature,* 303(5919), 696–698. https://doi.org/10.1038/303696a0

Borji, A. (2021). Saliency Prediction in the Deep Learning Era: Successes and Limitations. IEEE *Transactions on Pattern Analysis and Machine Intelligence,* 43(2), 679–700. https://doi.org/10.1109/TPAMI.2019.2935715

Borji, A. (2019, May 24). *Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges.* http://arxiv.org/abs/1810.03716

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Bruce, N. D. B., & Tsotsos, J. K. (2005). Saliency Based on Information Maximization. In *NIPS'05:* Proceedings of the 18th International Conference on Neural Information Processing Systems (pp. 155–162).

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2017, April 6). *What do different evaluation metrics tell us about saliency models?* http://arxiv.org/abs/1604.03605

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology, 197*(3), 551–566. https://doi.org/10.1113/jphysiol.1968.sp008574

Cannon, M. W., & Fullenkamp, S. C. (1996). A model for inhibitory lateral interaction effects in perceived contrast. *Vision Research, 36*(8), 1115–1125. https://doi.org/10.1016/0042-6989(95)00180-8

Dahou Djilali, Y. A., McGuinness, K., & O'Connor, N. (2024). Learning Saliency From Fixations. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 382–392). https://doi.org/10.1109/WACV57701.2024.00045

Ding, G., İmamoğlu, N., Caglayan, A., Murakawa, M., & Nakamura, R. (2022). SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing,* 120, 104395. https://doi.org/10.1016/j.imavis.2022.104395

Droste, R., Jiao, J., Noble, J.A. (2020). Unified Image and Video Saliency Modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12350. https://doi.org/10.1007/978-3-030-58558-7_25

Engel, F. L. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research, 11*(6), 563–576. https://doi.org/10.1016/0042-6989(71)90077-0

Engel, F. L. (1974). Visual conspicuity and selective background interference in eccentric vision. *Vision Research*, 14(7), 459–471. https://doi.org/10.1016/0042-6989(74)90034-0

Engel, S., Zhang, X., & Wandell, B. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature,* 388(6637), 68–71. https://doi.org/10.1038/40398

Eriksen, C. W., & Hoffman, J. E. (1972). Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics, 12*(2), 201–204. https://doi.org/10.3758/bf03212870

Falikman, M. V. (2015). *Struktura i dinamika zritel`nogo vnimaniya pri reshenii perceptivnyx zadach: konstruktivno-deyatel`nostny`j podxod*: dis. ... dokt. psixol. nauk: 19.00.01 (MSU).

Falikman, M. V., Utochkin, I. S., Markov, Yu. A., & Tyurina, N. A. (2019). Top-down regulation of visual search: does it exist in children? In *Cognitive Science in Moscow: New Research: Conference Proceedings,* Moscow, June 19, 2019 (pp. 513–517). BukiVedi. Institute of Practical Psychology and Psychoanalysis.

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biological Cybernetics, 46*(1), 27–39. https://doi.org/10.1007/BF00335349

Geiger, G., & Lettvin, J. Y. (1986). Enhancing the Perception of Form in Peripheral Vision. *Perception, 15*(2), 119–130. https://doi.org/10.1068/p150119

Gitman, Y., Erofeev, M., Vatolin, D., Bolshakov, A., & Fedorov, A. (2014). Semiautomatic visual-attention modeling and its application to video compression. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 1105–1109). https://doi.org/10.1109/ICIP.2014.7025220

Gorbunova, E. S. (2023). Mechanisms of representation construction in categorical search: the role of attention and working memory. *Russian Psychological Journal, 20*(3), 116–130. https://doi.org/10.21702/rpj.2023.3.6

Gusev, A. N., & Utochkin, I. S. (2012). The Influence of the Probability of Tips on the Efficiency of the Spatial Localization of the Visual Stimulus. *The bulletin of Irkutsk State University. Series «Psychology», 1*(1), 34–39. https://elibrary.ru/item.asp?id=18040183

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. In *Trends in Cognitive Sciences, 9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. https://doi.org/10.48550/arXiv.1512.03385

Helmholtz, H. von. (1896). *Handhuch der Physiologischen Optik (Zweite umgearbeitete Auflage).* Verlag von Leopold Voss.

Huang, G., Liu, Z., Maaten, L. van der, & Weinberger, K. Q. (2018, January 28). *Densely Connected Convolutional Networks.* https://doi.org/10.48550/arXiv.1608.06993

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. https://doi.org/10.1109/34.730558

Jampani, V., Ujjwal, Sivaswamy, J., & Vaidya, V. (2012). Assessment of computational visual attention models on medical images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing - ICVGIP '12*, (pp. 1–8). https://doi.org/10.1145/2425333.2425413

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2106–2113). https://doi.org/10.1109/ICCV.2009.5459462

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics, 54*(4-5), 245–251. https://doi.org/10.1007/BF00318420

Julesz, B. (1984). A brief outline of the texton theory of human vision. *Trends in Neurosciences, 7*(2), 41–45. https://doi.org/10.1016/s0166-2236(84)80275-1

Kachurka, V.A., Madani, K., Sabourin, C., Golovko, V.A., Kachurka, P.A. (2015). Object Detection in Computer Vision Systems: A Visual Saliency Based Approach. *Doklady BGUIR, 91*(5), 47–53.

Khokhlova, T.V. (2012). Current Views on Vision of Mammals. *Zhurnal obshchei biologii [Journal of General Biology], 73*(6), 418–434.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems,* 25.

Kruskop, A.S., Luniakova, E.G., Doubrovski, V.E., Garusev, A.V. (2023). Eye Movements in the Task of Visual Search Depending on the Verbalizability and Symmetry of Stimuli. *Lomonosov Psychology Journal*, 46 (4), 88–111. https://doi.org/10.11621/LPJ-23-40

Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision, 22*(5), 7. https://doi.org/10.1167/jov.22.5.7

Kümmerer, M., Theis, L., & Bethge, M. (2015, April 9). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. http://arxiv.org/abs/1411.1045

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2022). Using the DeepGaze III model to decompose spatial and dynamic contributions to fixation placement over time. *Journal of Vision, 22*(14), 3964. https://doi.org/10.1167/jov.22.14.3964

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 4789–4798).

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). DeepGaze IIE: Calibrated Prediction in and Out-of-Domain for State-of-the-Art Saliency Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12919–12928).

Lou, J., Lin, H., Marshall, D., Saupe, D., & Liu, H. (2022). TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494, 455–467. https://doi.org/10.1016/j.neucom.2022.04.080

Lyudvichenko, V., Erofeev, M., Gitman, Y., & Vatolin, D. (2017). A semiautomatic saliency model and its application to video compression. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 403–410). https://doi.org/10.1109/ICCP.2017.8117038

Martynova, O.V., & Balaev, V. V. (2015). Age Related Changes in Functional Connectivity of The Resting State Networks. *Psychology. Journal of the Higher School of Economics, 12*(4), 33–47. https://psy-journal.hse.ru/2015-12-4/167444789.html

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. *International Journal of Computer Vision, 82*(3), 231–243. https://doi.org/10.1007/s11263-009-0215-3

McCallum, R. (1996). *Reinforcement learning with selective perception and hidden state: PhD thesis.* University of Rochester.

Medioni, G., & Mordohai, P. (2005). Saliency in Computer Vision. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 583–585). Academic Press. https://doi.org/10.1016/B978-012375731-9/50099-9

Milanese, R. (1993). *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation: PhD thesis*. Univ. Geneva.

Minski, M. & Pejpert, S. (1971). *Perseptrony.* Mir.

Podladchikova, L. N., Samarin, A. I., Shaposhnikov, D. G., Koltunova, T. I., Petrushan, M. V. & Lomakina, O. V. (2017). *Sovremenny`e predstavleniya o mexanizmax zritel`nogo vnimaniya.*

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance* (Vol. 10, pp. 531–556). Erlbaum.

Posner, M. I., Cohen, Y., & Rafal, R. D. (1982). Neural systems control of spatial orienting. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 298*(1089), 187–198. https://doi.org/10.1098/rstb.1982.0081

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. Annual Review of Neuroscience, 35, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research, 42*(11), 1447–1463. https://doi.org/10.1016/s0042-6989(02)00040-8

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295x.85.2.59

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Remington, R., & Pierce, L. (1984). Moving attention: Evidence for time-invariant shifts of visual selective attention. *Perception & Psychophysics, 35*(4), 393–399. https://doi.org/10.3758/bf03206344

Rozhkova, G.I., Belokopytov, A.V., Iomdina, E.N. (2019). Present View of The Human Peripheral Vision Specifics. *Sensory Systems Journal, 33*(4), 305–330. https://doi.org/10.1134/S0235009219040073

Rubtsova, O. S., & Gorbunova, E. S. (2022). The Manifestation of Incidental Findings in Different Experimental Visual Search Paradigms. Psychology in Russia: State of the Art, 15(4), 140–158. doi: 10.11621/pir.2022.0409

Saarinen, J., & Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proceedings of the National Academy of Sciences of the United States of America, 88*(5), 1812–1814. https://doi.org/10.1073/pnas.88.5.1812

Sapronov, F.A., Gorbunova, E.S. (2025). Comparing AI-gene-rated stimuli and photos: visual search study. *Lomonosov Psychology Journal,* 48 (2), 109–131. https://doi.org/10.11621/LPJ-25-14

Sapronov F.A., Makarov I.M., Gorbunova E.S. Categorization in Hybrid Search: A Study Using Eye Movement Registration. *Eksperimental'naya psikhologiya = Experimental Psychology (Russia)*, 2023. Vol. 16, no. 3, pp. 121–138. https://doi.org/10.17759/exppsy.2023160308

Sechenov, I. M. (1942). Refleksy golovnogo mozga. AN SSSR.

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences, 10*(1), 38–45. https://doi.org/10.1016/j.tics.2005.11.008

Shevel, T. M., & Falikman, M. V. (2022). "Gaze cues" as a key to the mechanisms of joint attention: Main research results. *Cultural-Historical Psychology, 18*(1), 6–16. https://doi.org/10.17759/chp.2022180101

Sholle, F. (2023). *Glubokoe obuchenie na Python (2-nd edition).* Piter.

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition.* http://arxiv.org/abs/1409.1556

Sun, X., Houssin, R., Renaud, J., & Gardoni, M. (2019). A review of methodologies for integrating human factors and ergonomics in engineering design. *International Journal of Production Research, 57*(15–16), 4961–4976. https://doi.org/10.1080/00207543.2018.1492161

Utochkin, I. S., & Falikman, M. V. (2006). Inhibition of Return. Part I. Kinds and Properties. *Psychological Journal, 27*(3), 42–48. https://elibrary.ru/item.asp?id=9212401

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. https://doi.org/10.48550/arXiv.1905.11946

Theeuwes, J. (2013). Feature-based attention: It is all bottom-up priming. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1628), 20130055. https://doi.org/10.1098/rstb.2013.0055

Treisman, A. M. (1982). Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology. *Human Perception and Performance, 8*(2), 194–214. https://doi.org/10.1037//0096-1523.8.2.194

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1–2), 507–545. https://doi.org/10.1016/0004-3702(95)00025-9

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. https://doi.org/10.1037/0033-295x.84.4.327

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Voronin, I.A., Zakharov, I.M., Tabueva, A.O., & Merzon L.A. (2020). Diffuse Decision-Making Model: Assessment of The Speed and Accuracy of Answers in The Problems of Choosing from Two Alternatives in The Study of Cognitive Processes And Abilities. *The Theoretical and Experimental Psychology, 13*(2), 6–24. https://www.elibrary.ru/item.asp?id=48627007

Wang, W., Shen, J., Xie, J., Cheng, M.-M., Ling, H., & Borji, A. (2021). Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 220–237. https://doi.org/10.1109/TPAMI.2019.2924417

Wilson, H. R., & Bergen, J. R. (1979). A four mechanism model for threshold spatial vision. *Vision Research, 19*(1), 19–32. https://doi.org/10.1016/0042-6989(79)90117-2

Wolfe, J. M. (2012). Saved by a Log: How Do Humans Perform Hybrid Visual and Memory Search? *Psychological Science, 23*(7), 698–703. https://doi.org/10.1177/0956797612443968

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review, 28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Cave, K. R., & Franzel, 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. https://doi.org/10.1037//0096-1523.15.3.419

Yarbus, A. L. (1965). *Rol dvizhenij glaz v processe zreniya.* Nauka.

Received: July 5, 2025
Revised: September 20, 2025
Accepted: September 21, 2025

Author Details

Denis V. Yavna — Cand.Sci (Psychology), Associate Professor, Southern Federal University, Rostov-on-Don, Russian Federation; Scopus Author ID: 56034231500; WoS Researcher ID: B-1314-2013; ORCID ID: https://orcid.org/0000-0003-2895-5119; e-mail: dvyavna@sfedu.ru

Conflict of Interest Information

The authors have no conflicts of interest to declare.