NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Sofia Krasovskaya, Georgiy Zhulikov,

Joseph MacInnes

DEEP LEARNING NEURAL NETWORKS AS A MODEL OF SACCADIC GENERATION

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: PSYCHOLOGY
WP BRP 93/PSY/2018

SERIES: *PSYCHOLOGY*

Sofia Krasovskaya[1,2], Georgiy Zhulikov[1], Joseph MacInnes[1,2]

DEEP LEARNING NEURAL NETWORKS AS A MODEL OF SACCADIC GENERATION[3]

Approximately twenty years ago, Laurent Itti and Christof Koch created a model of saliency in visual attention in an attempt to recreate the work of biological pyramidal neurons by mimicking neurons with centre-surround receptive fields. The Saliency Model has launched many studies that contributed to the understanding of layers of vision and the sphere of visual attention. The aim of the current study is to improve this model by using an artificial neural network that generates saccades similar to how humans make saccadic eye movements. The proposed model uses a Leaky Integrate-and-Fire layer for temporal predictions, and replaces parallel feature maps with a deep learning neural network in order to create a generative model that is precise for both spatial and temporal predictions. Our deep neural network was able to predict eye movements based on unsupervised learning from raw image input, as well as supervised learning from fixation maps retrieved during an eye-tracking experiment conducted with 35 participants at later stages in order to train a 2D softmax layer. The results imply that it is possible to match the spatial and temporal distributions of the model to spatial and temporal human distributions.

JEL Classification: Z

Key words: saccade generation, salience model, deep learning neural network, visual search, leaky integrate and fire

[1] Vision Modelling Laboratory, Faculty of Social Science, National Research University Higher School of Economics, Moscow, Russia

[2] School of Psychology, National Research University Higher School of Economics, Moscow, Russia

[3] This Working Paper is an output of the research project “Interdisciplinary studies of vision” implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Introduction

This study is dedicated to research in the field of artificial neural network models of human vision in an attempt to understand human visual processes with neural networks. The goal was to create a neural network that would model layers of the human superior colliculus, with most of the focus concentrated on the way humans choose and generate saccades. The aim of the study was to obtain a deeper insight into the biological mechanisms of human saccadic production and to develop a foundation for a biologically plausible model in the future.

Previous attempts to recreate the work of the human visual system include the seminal salience model of Laurent Itti and Christof Koch [8]. Since then, one important question that has arisen is whether the biological accuracy of older models or newer approaches are more efficient in the recreation of the visual system.

Modelling in human vision

Contemporary cognitive science may be viewed as a product of the cognitive revolution, which embraced a tendency to use an interdisciplinary approach in research in order to gain a better understanding of cognitive processes. One of the main methods in this approach resembles reverse engineering, which is based on the notion that through the creation and investigation of artificial models of mental processes it is possible grasp the complex notion of human cognition in a more effective way. Based on this approach, it is possible to apply existing tools to knowledge in order to address human cognition holistically. One of the applications of reverse engineering in cognitive psychology includes the combination of computational modelling in human vision research.

Some of the best approaches to computational modelling focus on both the biological and theoretical aspects of vision, such as the classical Itti and Koch salience model (Figure 1) [8]. This model combines the biological aspect, describing the way that pyramidal cells work in the receptive fields of the visual cortex, and the feature integration theory, which suggests that focused attention is comprised of multiple object feature recognition and integration [19]. Such models were novel in their time in terms of spatial distribution accuracy, but perform poorly in terms of temporal distributions [13].

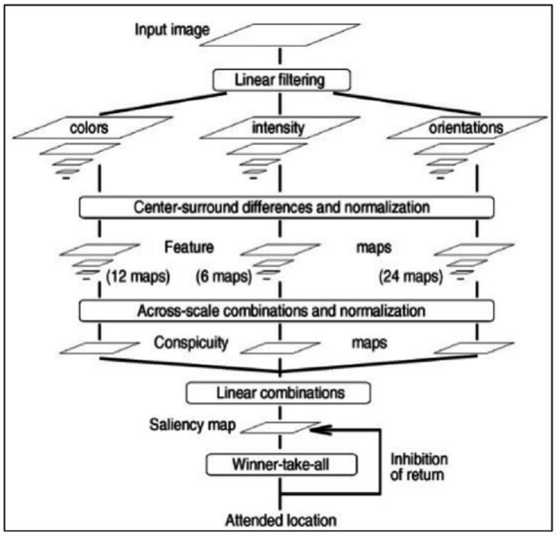

Figure 1. A schematic representation of Itti and Koch's [8] salience model.

In addition to models of spatial salience, temporal models predict the time course distribution of human saccades. The classic Itti and Koch integrated temporal prediction in addition to image processing, but more recent solutions do not. The most prominent examples of such models are the 'Leaky integrate-and-fire' (LIF) or 'spiking' network [12], [20] and drift diffusion algorithms [16], [17]. The leaky algorithm, otherwise called a spiking network, is similar to a biological neural network. Just like biological neurons fire upon reaching a threshold, the spiking network uses accumulation of evidence to gain a certain amount of information from an incoming signal and pass this information once it receives enough input. A distinctive feature of such neurons is that the accumulated data 'leaks' when the input signal ceases.

In general, each approach has its advantages and disadvantages based on the field of application and the potential tasks. However, in order to model human saccadic eye movements it is important to have a balanced combination of all the advantages of the existing models, which supposes good temporal and spatial characteristics, as well as high biological plausibility. Consequently, it is crucial to find the golden mean for modelling saccadic eye movements. In an attempt to find this golden mean, deep learning neural networks have been chosen as an instrument to model and investigate the visual system.

Deep Learning: a tool for modelling cognitive processes

Deep learning algorithms have proved to be useful in different spheres, such as biology, chemistry, finances, business, physics and neuroscience, as well as in many other fields of study [1]. These algorithms are a versatile, accurate and powerful tool, which has been acknowledged and proved by recent studies in object and speech recognition, as well as in other domains linked to complex data analysis [10].

In the field of visual research and eye-movement prediction, deep neural networks have become the leaders among other approaches. For instance, there prevails a tendency to choose more complex deep hierarchical structures over the Itti and Koch [8] salience model [4], [9]. The reason for this is due to the fact that deep learning algorithms are effective tools for modelling high levels of abstraction [2], such as vision and speech processes. Artificial models tend to have difficulties predicting fixations based on semantic content. People tend to look at points in space that are meaningful in a specific context.   The fact that models are bad at such semantic-related predictions is called the semantic gap. Deep learning models also provide a possibility to model human visual attention more accurately from the point of view of biological plausibility by narrowing the semantic gap between model and human predictions [7].

Modelling in human vision

The main purpose of the study is to gain a deeper understanding of how the human superior colliculus generates eye movements with the help of computational modelling.

The problem of most existing models, such as the Itti and Koch salience model [8], is that they focus primarily on the spatial accuracy of the predictions [4]. However, in terms of temporal accuracy, the reaction times of such models are too fast as compared to human reactions, which leads to poor matches of temporal distributions in such models to existing human distributions. LIF used in the models also produces bimodal RT distributions where human distributions tend to be log normal (Figure 2) [13], [14]. Based solely on temporal accuracy, there are models that better match human reaction times, such as diffusion models [13], or leaky neuron models [20], otherwise called a spiking model [5], but these models tend to abstract space to achieve that temporal performance.

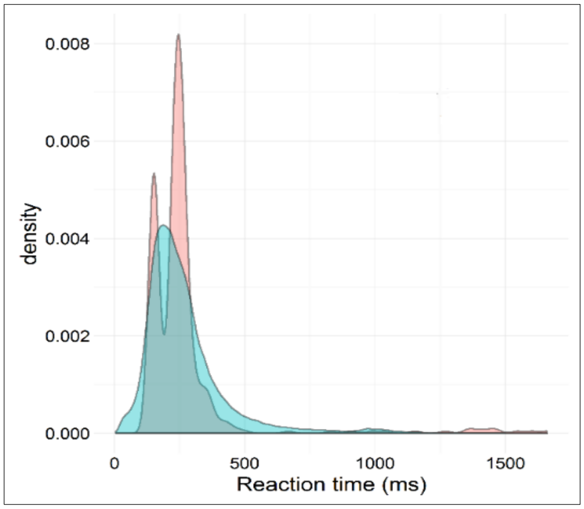

Figure 2. A demonstration of the bad matching using the default parameters of the Itti and Koch [8] salience model with a LIF layer to human saccade latency data [13]. 'Density' corresponds to the proportion of eye movements made in the time period.

Since the aim of this research focuses on the creation of an artificial neural network that would be able to learn to generate human-like, it was important to choose an algorithm that would mimic this generative nature of the visual system. For this it was necessary to take into account both the spatial and the temporal aspects of the process. Therefore, a combination of deep-learning algorithms based on salience maps and a LIF component were chosen as the main aspects of the model. Despite its limitations, the LIF algorithm was chosen over diffusion [13], [17], accumulator [20] and race [15], [22], [23] models. LIF models have accurate 2D spatial representations of visual space and are still best for natural scene processing.

Considering the methods used in the study, there were several potential approaches to the creation of a saliency map generator (Table 1). All approaches are similar in the way they separated the spatial and temporal component to focus on improved salience map generation.

Table 1. The general parameters of the approach.

Input	Real images (feed forward) + human fixation maps (feedback for softmax)
Supervised	Yes (early layers remained unsupervised)
Filters required	No
Toolbox	Matlab Neural Network Toolbox; Saliency Toolbox
Spatial component algorithm	Autoencoders + Softmax
Temporal component algorithm	Leaky integrate-and-fire

The approach implied using the Matlab Neural Network toolbox (https://www.mathworks.com/products/neural-network.html) and proposing a combination of autoencoders and a softmax layer. This model included two layers of autoencoders for unsupervised learning preceding a softmax layer for supervised learning. The dataset consisted of 59 training images and 18 testing images with dimensions of 1024x768 with testing and training fixation maps with dimensions 33x50 per image for each of 35 participants, which made a total of 1115 training maps and 478 testing maps.

Step 1. Each image was proportionally rescaled to a width of 120 and a height of 90 and reshaped from a matrix into a vector array of values between 0-255. Due to the fact that many salient points never reached figures higher than 100, the values were normalized to a distribution of values between 0 and 1 in order for these points to be vivid on the map. Each image was paired with matching fixation maps and their order was randomised.

Step 2. The processed training images were used as input to the first autoencoder layer of the model. The parameters of this layer were set to 200 iterations and hidden layers to size of 50. The learned hidden features were then extracted and saved as a separate variable.

Step 3. The extracted features from layer 1 were passed on to the second autoencoder layer for further training. The parameters of this layer were set to 200 iterations and a hidden size of 50. The learned hidden features were then extracted and saved as a separate variable.

Step 4. The extracted features from the second layer together with the processed training fixation maps were used to train the softmax layer. The parameters were set to default with 100 iterations.

Step 5. The autoencoder layers and the softmax layer were stacked together to form a deep network. The stacked network was trained on the training images and training maps. The trained deep network with all its parameters and values was then saved as a separate file for convenience during the testing phase.

Step 6. The network was tested on the training data. The predictions were reshaped to a dimension of 33x50 and normalised. This step was important in order to normalise and smoothen all matrix values between 0 and 1, otherwise the highest and lowest values were unevenly spread between 0 and 255, with the lower values tending to zero and the higher values tending to 255, which lead to a highly contrasted map.

Step 7. The network was tested on the testing data. The predictions were reshaped to a dimension of 33x50 and normalised using the abovementioned formula.

Step 8. For visual assessment, resulting salience maps were plotted in pairs of training maps against predictions based on the training data and testing maps against predictions based on the testing data.

Step 9. Before passing the images to the Leaky integrate-and-fire layer, they were processed with regards to the input requirements of the algorithm. They were rescaled proportionally to a width of 120 and a height of 90, transformed to grayscale. The predictions of the deep network with values normalised between 0 and 1 were reshaped from a matrix to an array.

Step 10. All 77 images together with the predicted fixation maps were passed on to the Leaky integrate-and-fire layer created in the Saliency Toolbox [21] which uses a ‘winner take all’ algorithm to produce a reaction time. The temporal limit was set to 2 seconds and the parameters were set to default and are listed in table 2:

Table 2. Parameters of the LIF model.

Parameter	**Description**	**Value**
timeStep	time step for integration ($1*10^{-3}$ ms)	0.001 seconds
Eleak	leak potential	0 volts
Eexc	potential for excitatory channels	100e-3 volts
Einh	potential for inhibitory channels	-20e-3 volts
Gleak	leak conductivity	1e-8 Siemens
Gexc	conductivity of excitatory channels	0 Siemens
Ginh	conductivity of inhibitory channels	0 Siemens
GinhDecay	time constant for decay of inhibitory conductivity	1 Siemens
Ginput	input conductivity	5e-8 Siemens
Vthresh	threshold potential for firing	0.001 volts
C	capacity	1e-9 farads
V	current membrane potential	0 volts
I	current input current	0 amperes

Results

The approach proved to be efficient in terms of spatial accuracy. The model predicted fixation maps similar to human fixation maps. As seen in figure 4, the model produced maps based on all the human fixation maps per image.

The current results demonstrate that the deep-learning model based on the Itti and Koch saliency model [8] is effective in generating salience maps of visual space.

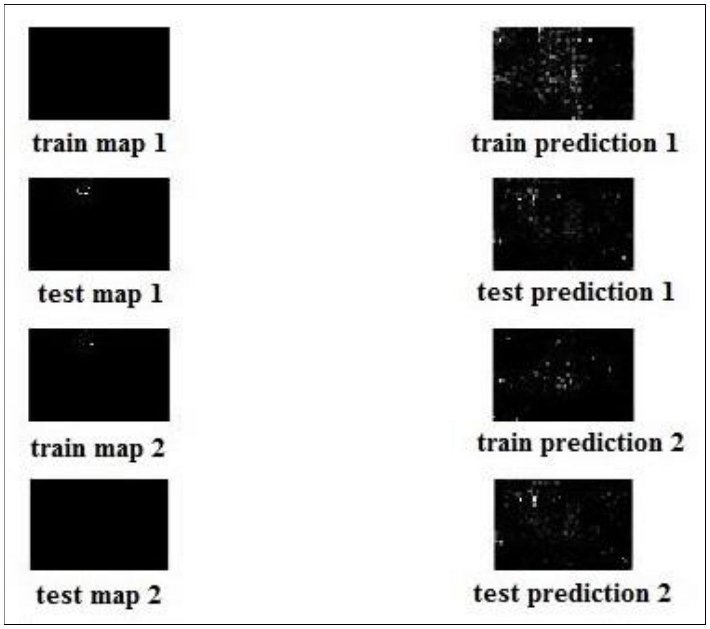

Figure 4. A demonstration of the autoencoder+softmax model spatial predictions.

In terms of temporal accuracy, however, the leaky integrate-and-fire algorithm did not entirely meet our expectations. Despite addressing, theoretically, both spatial and temporal issues of modelling saccadic eye movements, the classic LIF approach with default parameters has proven to have limitations in terms of temporal biological plausibility. Results have shown good findings with regard to the mean and z-test, however, the overall distribution has proved the approach to be ineffective regarding biological accuracy [14]. The reaction times produced by the model fit into the range of human saccadic reaction times, however, they seem to have a deterministic nature when used with default parameters. Besides, the bimodality in the distribution was present throughout all testing phases (figure 5).

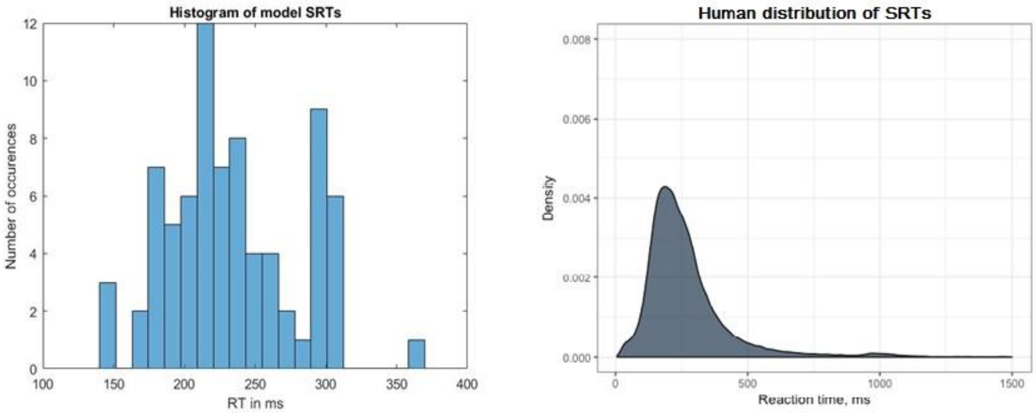

Figure 5. Distribution of SRTs produced by the model (left) against human distribution (right).

Metrics for result analysis

In order to evaluate the performance of the saliency algorithm it was necessary to implement certain evaluation metrics. As the saliency part of the network dealt with spatial locations, the reasonable choice was to use a location-based metric algorithm in order to assess the performance of the generated saliency maps. One characteristic instrument of such metrics is the area under the Receiver Operating Characteristic curve (AUC ROC), which estimates the tradeoff between true positive and false positive values at different discrimination thresholds by verifying that true positives are labelled before negative values [6], [4]. The AUC unit of measurement is one of the most widely used instruments for saliency assessment [9], as shown on the MIT saliency benchmark website [3].

The specific AUC metric that was chosen for the evaluation task was the AUC-Judd [9], [18]. The AUC-Judd interprets fixations as a classification task, where a pixel of the map may be either salient or not by applying a threshold over the intensity value of the saliency map[11]. Each salient pixel against human fixations on the map is considered a true positive value, whereas salient pixels over non-fixation areas are classified as false positive values. The final AUC score is then calculated and plotted as a tradeoff between true and false positive values. The highest possible score may be 1, whereas a 0.5 score is considered as random performance. Thus, any score lower than 0.5 is regarded as unsatisfactory.

Our proposed saliency model has proven to have an above average performance on the AUC-Judd test. The average score on the test for 77 generated maps was estimated as 0.60, which is equal to the score of the first model proposed by Dirk Walther and Christof Koch for the MIT saliency benchmark website [21], with the lowest score among the maps equal to 0.48 (figure

6) and the highest score equal to 0.70 (figure 7).

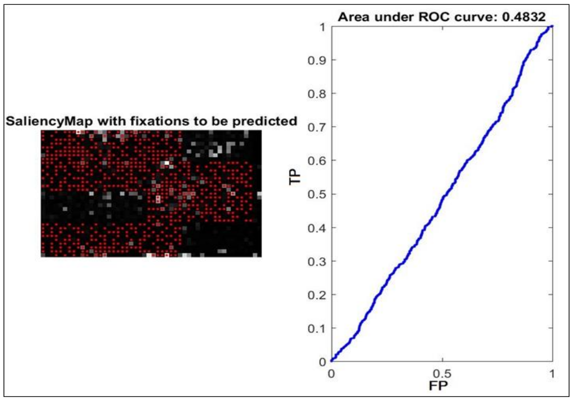

Figure 6. The minimal score of the generated maps on the AUC-Judd test. FP are the False Positive variables, TP – the true positive variables.

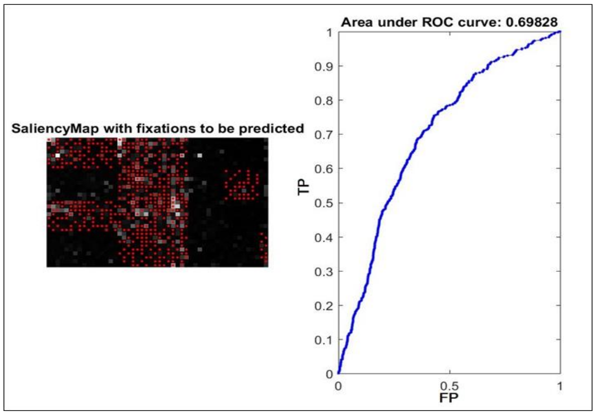

Figure 7. The maximal score of the generated maps on the AUC-Judd test. FP are the False Positive variables, TP – the true positive variables.

Conclusion

The current study has demonstrated that it is possible to improve classical theories with the help of new tools, such as deep learning neural networks. Just like Itti & Koch’s salience model [8] showed how pyramidal neurons work, the described model is an attempt to partially visualize how the superior colliculus works by using ‘accumulation of evidence’ modelling. Each layer of the network makes different contributions to the model, just like the human visual system processes visual information, which provides a good foundation for future research. However, it is necessary to further investigate the temporal aspect as well as to focus on the frames of reference in the human visual system in order to improve spatial predictions of the model in a physiologically accurate manner.

References

[1] Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods, 43(1), 3-31.

[2] Bengio, Y. (2009). Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2(1), 1-127.

[3] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). Mit saliency benchmark.

[4] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models?, 1–24. Retrieved from http://arxiv.org/abs/1604.03605

[5] Das, S., & Pedroni, B. (n.d.). Implementation of a Restricted Boltzmann Machine in a Spiking Neural Network. Isn.Ucsd.Edu. Retrieved from http://www.isn.ucsd.edu/courses/bggn260/2012/reports/Das_Pedroni.pdf

[6] Ferri, C., Hernández-Orallo, J., & Flach, P. A. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 657-664).

[7] Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 262-270).

[8] Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40(10–12), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

[9] Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations.

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[11] Kachurka, V., Madani, K., Sabourin, C., & Golovko, V. (2015, June). From human eye fixation to human-like autonomous artificial vision. In International Work-Conference on Artificial Neural Networks (pp. 171-184). Springer, Cham.

[12] Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. Neural Networks, 10(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

[13] MacInnes, W. J. (2017). Comparison of temporal models for spatial cuing. 5th annual Polish Eye tracking conference

[14] Merzon, L., Malevich, T., Krasovskaya, S., Zhulikov, G., Kulikova, A., MacInnes, W. J. (2018). Temporal Limitations of the Standard Leaky Integrate and Fire Model. Manuscript submitted for publication.

[15] Purcell, B. A., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2012). From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. Journal of Neuroscience, 32(10), 3433-3446.

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

[17] Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for

Two-Choice Decision Tasks. Neural Computation, 20(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

[18] Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013, December). Saliency and human fixations: state-of-the-art and study of comparison metrics. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 1153-1160). IEEE.

[19] 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

[20] Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

[21] Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. Neural networks, 19(9), 1395-1407.

[22] Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. Psychonomic Bulletin & Review, 1(2), 202–238. https://doi.org/10.3758/BF03200774

[23] Wolfe, J. M., & Gancarz, G. (1997). Guided Search 3.0. In Basic and clinical applications of vision science (pp. 189-192). Springer, Dordrecht.

Corresponding author
Sofia Krasovskaya
PhD student at School of Psychology, National Research University Higher School of Economics.
Research assistant at Vision Modelling Laboratory at Faculty of Social Science, National Research University Higher School of Economics
E-mail: krasov.sofia@gmail.com

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.