NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

*Liya Merzon, Georgiy Zhulikov, Tatiana Malevich,*
*Sofia Krasovskaya, Joseph MacInnes*

# TEMPORAL LIMITATIONS OF THE STANDARD LEAKY INTEGRATE AND FIRE MODEL

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: PSYCHOLOGY
WP BRP 94/PSY/2018

This Working Paper is an output of a research project 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

***SERIES:*** *PSYCHOLOGY*

*Liya Merzon[1,23], Georgiy Zhulikov[1], Tatiana Malevich[1,4], Sofia Krasovskaya[1,2], Joseph MacInnes[1,2]*

# TEMPORAL LIMITATIONS OF THE STANDARD LEAKY INTEGRATE AND FIRE MODEL[5]

The leaky integrate and fire model of neural spiking has been used extensively to simulate saccadic responses in a variety of tasks from visual search to simple reaction times. Although it has been tested for its neural spiking accuracy and its spatial prediction of fixations in visual salience, it has not been well tested for its temporal accuracy. Saccade generation invariably results in a positively skewed distribution of saccadic reaction times over large numbers of samples, yet we show that the LIF algorithm tends to produce a distribution shifted to shorter fixations (in comparison with human data) in its classic implementation. Further, parameter optimization using a genetic algorithm and Nelder–Mead method does improve the fit of the resulting distribution, but is still unable to match temporal distributions of human responses in a simple visual search task. Further analysis revealed, that the LIF algorithm produces discrete reaction times instead of distributions. Aggregated over many pictures they may be treated as a distribution although the form of this distribution depends on the input images used to create it.

[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] Center for Cognition & Decision Making, National Research University Higher School of Economics, Moscow, Russia
[4] Werner Reichardt Centre for Integrative Neuroscience, University of Tuebingen, Tuebingen, Germany
[5] 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.

JEL Classification: Z

Key words: saccade generation, salience model, visual search, leaky integrate and fire model

# Introduction

Despite limits to the processing capacity of the human visual system, we are quick to make sensible interpretations of incoming visual information in real time. This ability to select information from our complex environment is commonly ascribed to attention, the focus of which is likened to a spotlight moving across the visual field that highlights its most relevant areas (Carrasco, 2011). Shifts of attention can be endogenous, i.e. top-down and goal-directed, or exogenous, i.e. bottom-up and driven by external factors such as perceptual properties of visual stimuli (Posner, 1980; Carrasco, 2011).

Dominant models of bottom-up attention and visual search rely on the idea that visual saliency influences where we attend, i.e. properties of a visual stimulus stand out against other environmental stimuli and capture our attention (Koch & Ullman, 1985; Itti & Koch, 2000; Walther & Koch, 2006; see Itti & Borji, 2013 for a review). This concept is based on the feature integration theory of attention (Treisman & Gelade, 1980; also see Itti & Borji, 2013), which states that at an early, 'preattentive' processing stage features are registered in parallel across the whole visual field and encoded along a number of perceptual dimensions (orientation, colour, spatial frequency, brightness, etc.). At a later, 'attentive' stage, these feature maps are normalized and integrated across spatial coordinates into a higher representation, i.e. a saliency map (SM), and the location of the most salient stimulus 'wins' a competition between neurons and is therefore attended. The final layers of this model for saccadic generation are typically implemented with a neuronal activation model in the leaky-integrate-and-fire (LIF) layer and spike on the winner-take-all (WTA) layer (Itti & Koch, 2001). Inhibition of return (IOR - Klein, 2000) is used as a mechanism to prevent refixations at the most salient locations and promote novelty in search (Klein & MacInnes, 1999).

The saliency model is thought to be biologically plausible, reflecting the center-surround receptive field interactions in visual pathways carried out in the pyramidal architecture of the model, feature-specific sensitivity of neurons in early visual cortex, as well as neuronal activation and spiking simulated in its LIF and WTA layers (Itti, Koch, & Niebur, 1998; Itti & Borji, 2013). Also, this family of models have been shown to produce a reasonable fit to human

data in terms of spatial localization of salient stimuli as measured by overt attention (fixations) (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002; Itti & Borji, 2013) (see also MIT saliency benchmark http://saliency.mit.edu/home.html for accuracy characteristics of spatial predictions by different implementations based on saliency model against human data).

The early parallel feature maps of the model combine to produce a salience map that provides location and intensity information, and this is used to generate the spatial predictions of fixation locations. The predictions for temporal distribution of these fixations, however, is generated by the LIF and WTA layer. These two-dimensional neural net layers simulate neuronal spikes with each pixel of the SM considered as a simplified version of a neuronal population, i.e. a neuron with powerful synaptic connections (Itti & Koch, 2000). These artificial neurons are built using differential equations to simulate a build-up of charge potential at a location and to fire a pulse once a threshold has been reached. The input current of the neurons is initially set to the value of saliency map (multiplied by a constant for scaling out range) and adding some noise. These LIF neurons make excellent predictions of single cell firing (Badel et al., 2008a, 2008b). Voltage of a neuron is updated iteratively according to the following formula and include parameters as listed in the Table 1:

$$V = V + \frac{dt}{C\cdot(I - Gleak\cdot(V - Eleak) - Gexc\cdot(V - Eexc) - Ginh\cdot(V - Einh))} \quad (1)$$

**Table 1. LIF parameters of Saliency Model (Walter & Koch, 2006). For more details about the specific implementation, see (Koch, 2004) and the Matlab SaliencyToolbox, http://www.saliencytoolbox.net implemented by Walhter and Koch (2006)**

| | |
|---|---|
| timeStep | time step for integration ($1*10^{-3}$ ms) |
| Eleak | leak potential |
| Eexc | potential for excitatory channels |
| Einh | potential for inhibitory channels |
| Gleak | leak conductivity |
| Gexc | conductivity of excitatory channels |
| Ginh | conductivity of inhibitory channels |
| GinhDecay | time constant for decay of inhibitory conductivity |
| Ginput | input conductivity |

| | |
|---|---|
| Vthresh | threshold potential for firing |
| C | capacity |
| V | current membrane potential |
| I | current input current |

After nearly 20 years, the standard salience model (Itti & Koch, 1999; Walther & Koch, 2006) is still used as a solid implementation of our theoretical understanding, however, other algorithms have surpassed it in fixation prediction and classification. Models based on deep convolutional networks, in particular, have surpassed other models in terms of spatial prediction accuracy. Indeed, according to MIT Saliency Benchmark (Bylinskii et al., 2015), the original Saliency Model, proposed by Itti and Koch, has moderate accuracy: AUC-Judd metric, a version of the Area Under ROC curve (Bylinskii et al., 2016), is equal to 0.6, while the best result in the Benchmark is reached by a model with a deep network whose accuracy is 0.88 (or 0.84 for the same model without including center bias in the model).

The second limitation, and the focus of this article, is found in the temporal predictions of saccadic generation. The original salience model maintained one advantage over the recent deep learning approaches in that it is was capable of generating human-like saccades in time as well as space. In the past few decades several approaches addressing temporal accuracy of responses have appeared (Ratcliff & McKoon, 2008; Purcell, 2012; Usher & McClelland, 2001), however, there is little research on how the classic model developed by Itti and Koch (2000) reproduces temporal dynamics of overt attention (overt attentional shift is related to voluntary or automatically eye movements towards the stimulus, in contrast with covert attention, which could be described as inner shift of attentional focus in a person's mind without direct relation to gaze movements; Posner, 1980). The latency of saccades in viewing or search task is measured as the time duration of the intervening fixation and the temporal distribution of saccades in search tend to have a distinct positively skewed distribution (Van Zandt, 2002; Wolfe, Palmer & Horowitz, 2010; MacInnes et al., 2014). In the present study, we propose to test the temporal accuracy of the classic LIF+WTA combination and determine which, if any, parameter space of the model allows for an accurate temporal fit of observed human data.

# Proposal

Given the Itti & Koch (2000) model's longevity and its ability to generate saccades using LIF, we wanted to test the model's accuracy in reproducing saccadic distributions against human data from a visual search task. Multiple attempts of optimal parameter choice were used to match human data, with the the initial attempt being the default parameter settings found in an implementation (Walther & Koch, 2006). To follow, we trained the LIF by adjusting its default parameters using a genetic algorithm (GA) (Davis, 1991) and a matlab build-in optimizer (fminsearch ), which uses the Nelder–Mead method (Nelder & Mead, 1965; MathWorks, 2018) to find the optimal parameter space for temporal prediction. While these parameters are also involved in spatial prediction (via the WTA component), our optimization only considered the temporal accuracy for this initial stage. If the LIF is able to simulate an accurate temporal distribution, then we could look further to see if a parameter space could be found where spatial and temporal accuracy could co-exist.

# Methods

To meet the goal, we modified the algorithm developed by Walther and Koch (2006) (the Matlab SaliencyToolbox, http://www.saliencytoolbox.net). This algorithm is an extended version of the Itti et al. (1998) implementation of the Koch and Ullman (1985) salience model that takes into account attending to proto-object regions (see Rensink, 2000) and incorporates feedback connections. The source toolbox code was modified by separating first stage salience map production from the LIF and WTA components of the model's implementation (Krasovskaya et al, 2018). Although the method of the model's prediction was not changed, this allowed testing temporal accuracy and modifying the LIF parameters separately from the spatial salience map. The inhibition of return component was disabled to exclude the additional influence and only the first fixation after the onset of the stimuli was used for each image.

The model's predictions were tested against human data collected from 45 test images presented on a 21" LCD monitor as a part of visual search experiment. For our purposes, we extracted the information about the timing of the first fixation made by participants after the onset of the presented image (the duration of the fixation that coincided with onset wasn't used, since it was influenced by the onset of the image). Fixations longer than 1500 ms were considered as outliers and were excluded from further analysis.

Human data were collected as a part of a previous study (Gordienko, 2016). Participants (N = 16) performed a visual search task on the images of natural indoor scenes taken from the LabelMe open database (Russell, Torralba, Murphy, & Freeman, 2008).

Figure 1. The examples of visual stimuli taken from the LabelMe database.

There were four blocks of 45 trials each, with two possible targets to search for, a "cup" and a "painting." The target remained the same within an experimental block but the order of images presented was randomized for each participant and the number of target objects varied for each image. Each trial began with the instruction to search for a particular target shown on the screen until the joystick button press. After that, a point of fixation (a cross) was displayed in the middle of the screen for one second. Participants were instructed to fixate at the cross. As the cross disappeared, the search image was presented for eight seconds; participants were asked to search through the image and specify the number of target objects shown after its removal by pressing the joystick up and down buttons. The figure below shows the trial sequence. Eye movements were monitored with the Eyelink 1000 eye-tracking system sampling at 1000 Hz.

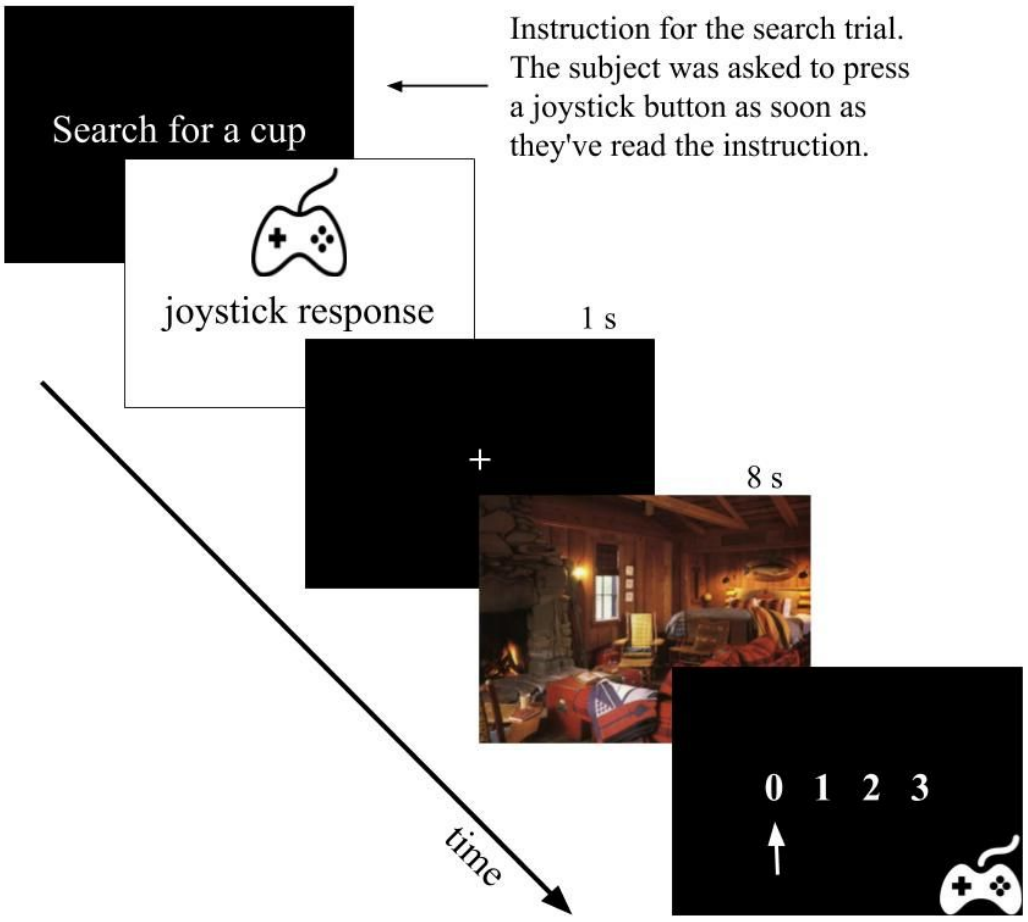


Figure 2. The trial sequence, rebuild by the description provided in Gordienko's paper (2016).

The current study used two dataset from the experiment described above. The first, smaller dataset, includes data, collected on 44 pictures. Information includes 782 first fixations (29528 fixations in total). This dataset was used in Experiments 1-3. The second, larger, dataset includes data, collected on 91 pictures: 1593 first fixations (60186 fixations in total); and was used in the Experiments 3-4.

# LIF parameters optimization

## Experiment 1. Check the default parameters

First, we tested the temporal predictions of the model with the default parameter space as defined in the article (Walther & Koch, 2006) against the first saccades, taken from human data of the first dataset.

## Results of Experiment 1.

The result is shown in the figure below:

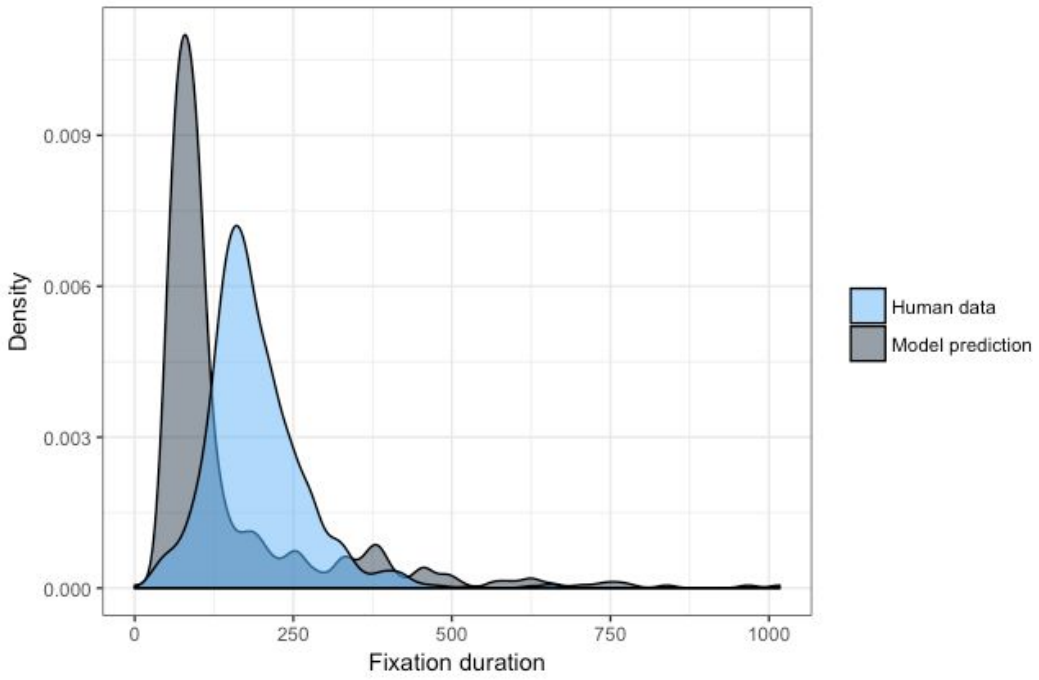


Figure 3. The predictions of the saliency model with the default parameters, suggested by Walther and Koch (2006) against the human data.

The default model produced faster fixation durations in comparison to the human data (mean = 150 ms against 190.8 ms) and showed a narrower standard deviation interval (*SD* = 146 against 77.7). We compared the generated distribution of times to those produced by human participants using the Kolmogorov-Smirnov test, and we were able to dismiss the null hypothesis that the two were sampled from the same distribution (ks-test D = 0.58067, p-value < 2.2e-16, see figure 3).

The default parameter space therefore can't be considered as acceptable for modeling temporal aspects of saccadic movements.

## Experiment 2. Parameter optimization

In order to obtain better temporal accuracy we optimized the default parameters of the saliency model with the genetic algorithm (GA) and Nelder-Mead method (NM). The optimized parameters included the parameter of the LIF (the potential for excitatory and inhibitory channels, the leak conductivity, the input conductivity, the threshold potential for firing, and capacity) and WTA (capacity, the leak conductivity, and the conductivity of inhibitory channels) layers as well as three noise parameters (the amplitude of random noise, the amplitude of constant noise, and the range of the saliency map output). Afterwards they were tested as part of the full WTA neural network.

For both optimization algorithms, the fitness function to minimize error used a combination of statistics from the Kolmogorov–Smirnov (KS) test and z-tests as compared to the observed human data distributions. The KS test fitness function was set to try to minimize the KS statistic when comparing the model and human data with

$$KS_{n,m} = sup\,|F_{1,n}(x) - F_{2,m}(x)|$$

where F1 and F2 are the distribution functions of the human and model RTs, and sup is the supremum function.

## Results of Experiment 2

The best result was obtained with the NM method. A z-test showed that the mean and standard deviation of two distributions were not statistically different (*z = 0.22378, p-value = 0.8229*), however, based on two-sample Kolmogorov-Smirnov test we were unable to dismiss the hypothesis that model and human data were from the same distribution (*D = 0.22506, p-value = 0.02939*). A visual inspection of the results (figure 4) also shows a multimodality in the generated data that is not typically observed in human data and doesn't show typical positive skewness.

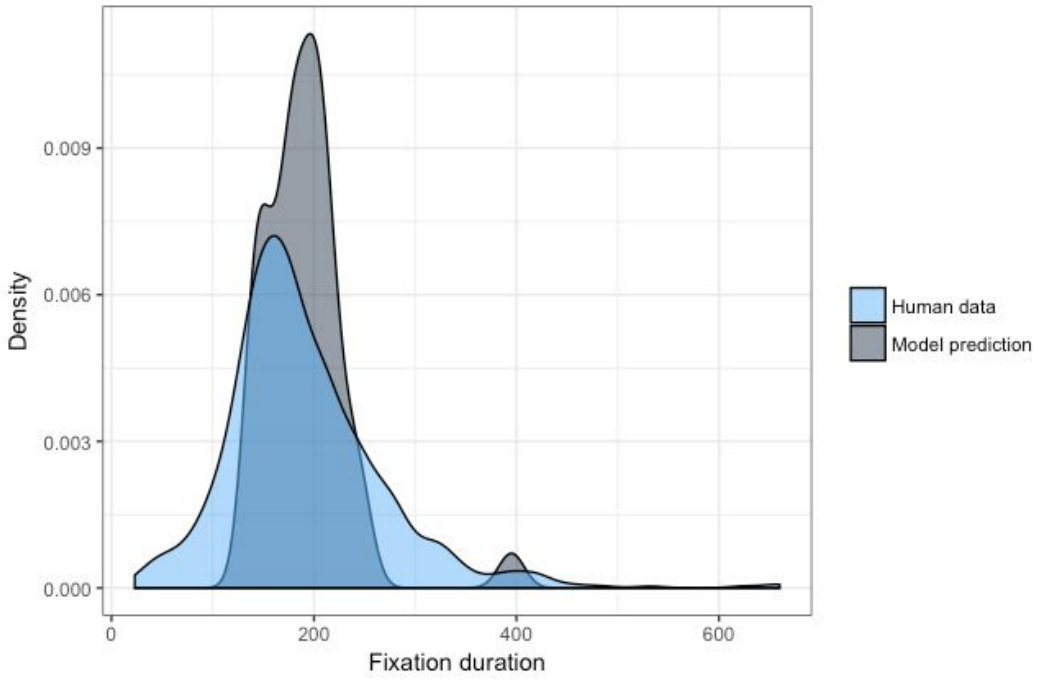


Figure 4. The predictions of the saliency model with the second set of parameters pictured against the human data. The model generated one initial saccade per image (for 44 pictures). Z-test statistics: z = 0.00075273, p-value = 0.9994. Ks-statistics D = 0.22506, p-value = 0.02939

Considering the result, we wanted to investigate if the capabilities of the saliency model allowed to improve the model's fit to ground truth. So, for the next experiment the model was set up to generate a bunch of initial saccades for each picture, which allowed to increase its flexibility.

## Experiment 3. Increased number of saccades

The saliency model has a random component (constant and random noise: parameters noiseAmpl and noiseConst of the Saliency Model) in the leaky integrate-and-fire layer that theoretically allows it to model distributions of reaction times. To give this parameter the best chance of influencing the resulting distribution, we increased the number of fixations that were generated by the model for each image. Fixations per image were increased to 10 although each of the 10 were generated as 'first' fixations independently so as to avoid IOR in the model. As a result, a distribution of 440 saccades (10 saccades for 44 images) were generated for comparison with the human data. In addition, we also increased our human dataset by adding data collected on other 47 images; there were 91 images in total in the new dataset, collected from the same 16 participants (see the "Methods" section for more details).

## Results of Experiment 3

The best parameters comparing the 10 generated saccades per image to the original 44 image dataset produced a distribution of the latencies of the initial fixations that was closer to the ground truth of human data. It matched with mean and standard deviation ($z$ = *-0.013418, p-value = 0.9893*), but for the KS-test the null hypothesis was rejected ($D$ = *0.19182, p-value = 2.007e-09*).

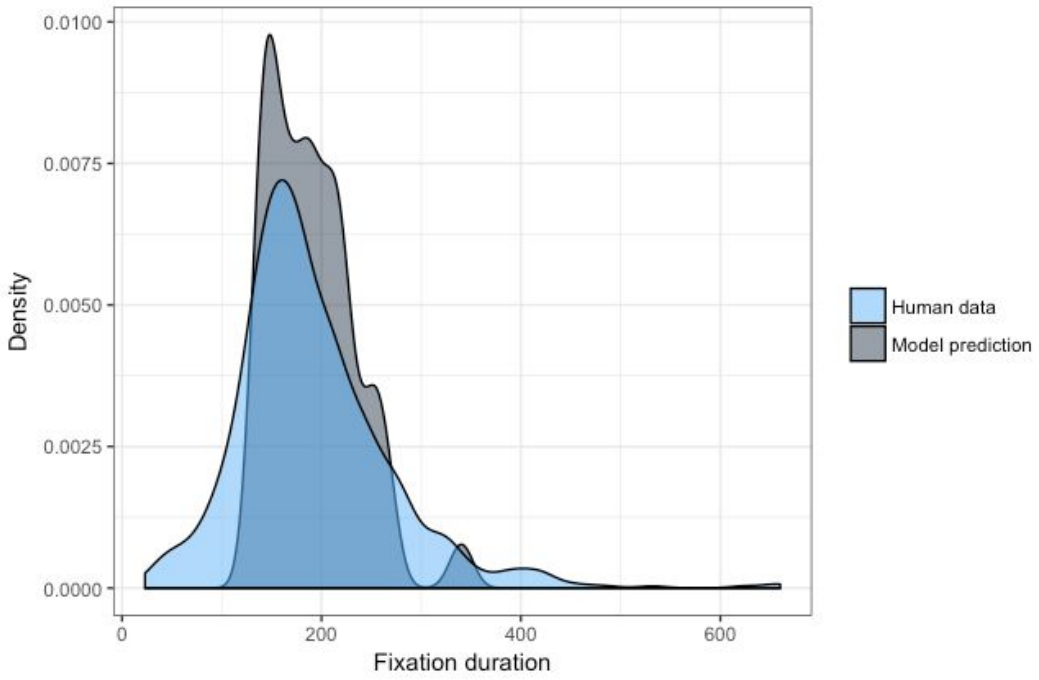


Figure 5. The saliency model's predictions with the third set of parameters pictured against the human data. The model generated ten initial saccade per image (for 44 pictures). Z-test statistics: z = -0.013418, p-value = 0.9893. Ks-statistics: D = 0.19182, p-value = 2.007e-09.

We repeated the test with additional data for the training dataset in order to allow the model learn better parameters. The new dataset contained information about 1593 initial saccades (60186 saccades totally), collected on 91 pictures in the same experiment, which is described above in the "Methods" section. The optimization procedure was the same used previously. Unexpectedly, the result, obtained on the larger dataset, was worse and didn't match even with mean of the distribution (*z = 3.1117, p-value = 0.00186*):

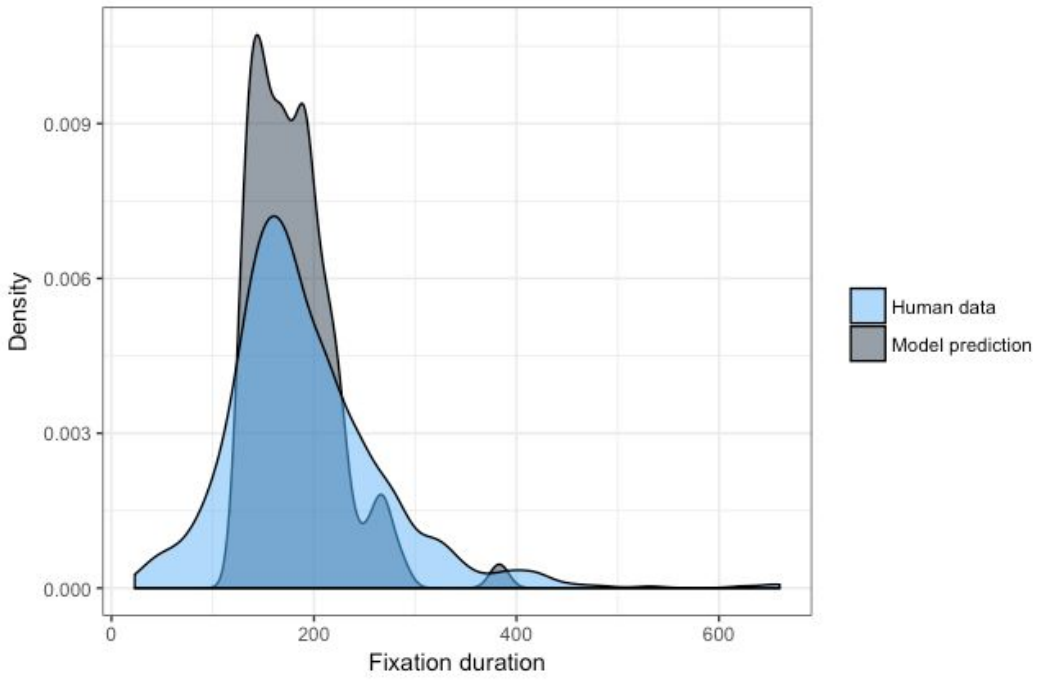


Figure 6. The saliency model's predictions with the fourth parameters space, learned on the 91 pictures dataset. The model generated ten initial saccade per image. Z-test statistics: z = 3.1117, p-value = 0.00186, Ks-statistics: D = 0.15857, p-value = 1.305e-09.

The analysis revealed that noise, which is supposed to increase similarity to the real human data, doesn't produce enough randomness in the model predictions. The default amplitude of random noise was $10^{-17}$, and the amplitude of constant noise was $10^{-14}$. This was apparently too small, and the model seemed to learn to produce a distribution of the saccadic latencies based only on differences in saliency of different images. In essence, the optimization algorithm learned to use the difference in image saliency to produce the distribution in responses. This, of course differed from the ground truth of human reactions.

## Experiment 4 Learning new parameters on data separated by pictures

As observed in human data, initial fixations for a single image also form a distribution by themselves. In our human dataset, there were 16-19 initial saccades for each image. We wanted to determine if the model could predict the duration of these initial fixations given a particular picture.

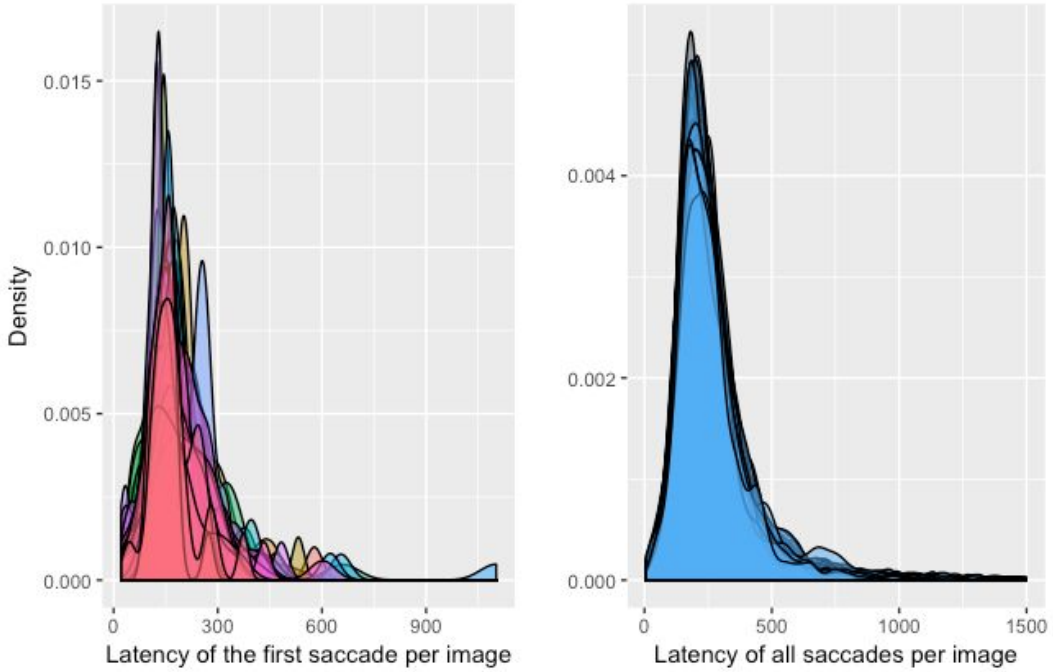


Figure 7. The distributions of saccades' latency for each image in the dataset. On the left: only initial saccades (16-19 data points for each image). On the right: all saccades (542-794 data points for each image, mean number 661). Both plots shows that latencies of saccades, which are related to only one image, nonetheless tend to produce a right-skewed distribution, typical for any reaction time distribution.

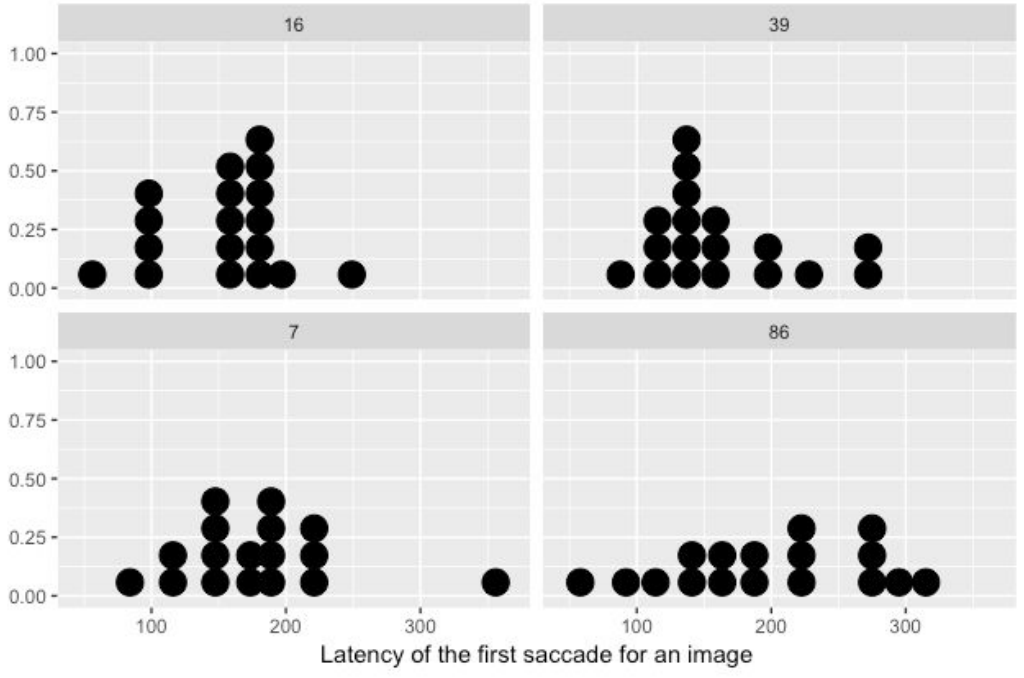


Figure 6. The dot plots of latencies of initial saccades for four random images in the reference human dataset (with the breaks equal to 20 ms).

The next version of the training algorithm was used to prevent our model from "cheating" and relying on different levels of input saliency to produce the distribution. The human data was divided into images, and an optimization function was set up as sum of ks statistics calculated for model predictions against ground truth for each image separately. Since we only had a limited number of first human fixations per image, we continued to use data from all fixations as the ground truth used to train the parameters for the image distributions.

## Results of Experiment 4

The best parameters highlighted the problems mentioned above, meaning that the algorithm was not able to find an accurate set of parameters when forced to incorporate variance from its own parameters rather than from the images. The distribution produced by the resulting parameter space was not close to the ground truth of human data (*ks-statistics D = 0.19327, p-value < 2.2e-16*; see the figure below).

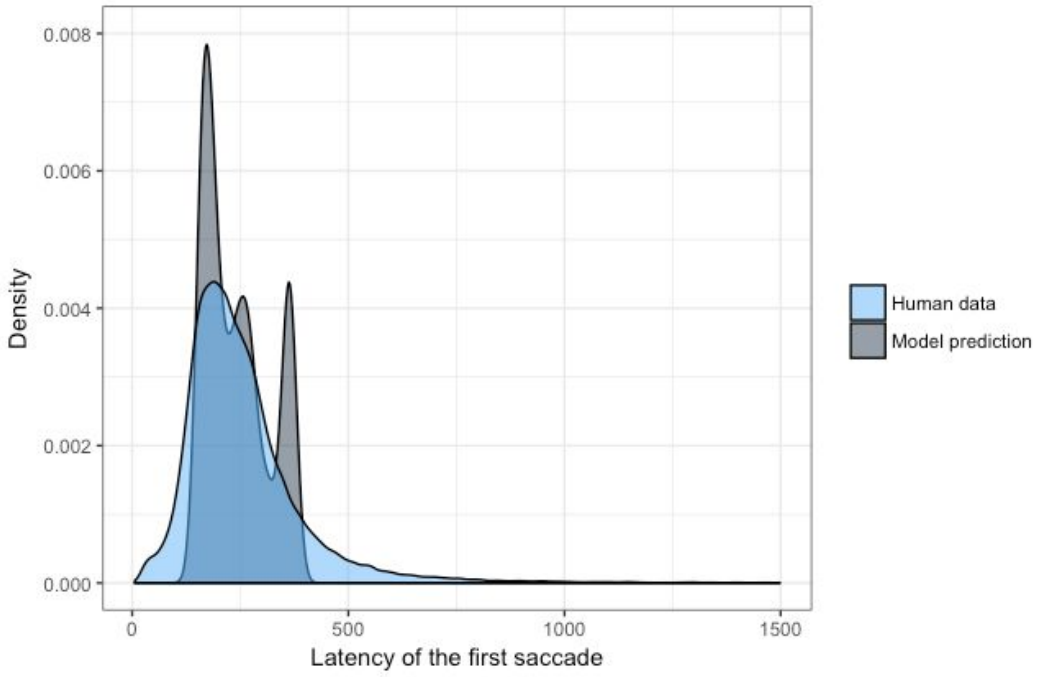


Figure 8. The saliency model's predictions with the fifth parameters space, learned on the 91 pictures dataset. The model generated ten initial saccade per image. Z-test statistics: z = 6.5611, p-value = 5.341e-11. Ks-statistics: D = 0.19327, p-value < 2.2e-16

The best noise constants were ($10^{-14}$ and $10^{-11}$ for random and constant noise respectively) and were able to produce only two different reaction times per image at most. An attempt to manually increasing the parameters, resulted in a model didn't generate a saccade within 2 sec interval. Manual decreasing the parameters led to only one fixation response per image. The

problem is that the noise added by these parameters is added to the saliency map directly which generally has features of the order $10^{-9}$. It does not impact the dynamics of the model, but only changing its initial values, which explains low variability between different runs on the same image.

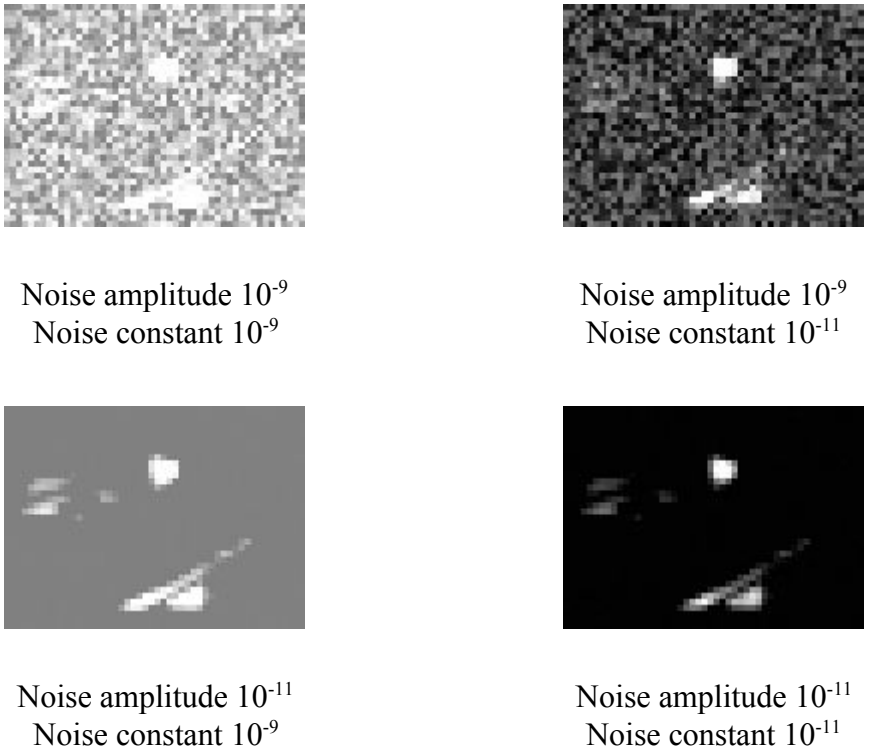


Figure 9. Examples of different noise parameters changing a saliency map

## Discussion

We have attempted to test the temporal accuracy of the classic LIF+WTA as part of the classic salience model in order to determine whether any parameter space of the model allows for an accurate temporal fit of observed human data.

The default parameters did not reproduce the temporal dynamics of human visual attention, nor did a model with optimized parameters. Further investigation showed that the result was highly dependent on variability in the images used as input to generate the fixations, with additional images even making the final distribution worse. The noise in LIF adds little change to the fixation durations, which makes them closer to constant values than to samples from a distribution.

Other models have been shown to simulate reaction times for visual search with high accuracy. One of such models is the drift diffusion model (MacInnes, 2017; Ratcliff & McKoon, 2008).

These models however restrict the problem by allowing only two signals to 'race' toward a single threshold. They also lack important theoretical features of the LIF such as lateral inhibition, which is characteristic of the biological visual system. Another model that shows good temporal resolution is the so-called 'race' model (Purcell, 2012; Wolfe, 1994, 2007). The main difference between race models and diffusion models is that race models have multiple signals competing for a single threshold, but, just like diffusion models, they do not involve lateral inhibition in the biological sense, nor do they have an accurate representation of retinal space. Finally, there are combinations of classic diffusion algorithms and LIF, the so-called 'leaky competing accumulator' models (Usher & McClelland, 2001), which include biologically important features, e.g. an accumulation drop during signal loss (leaky) and lateral inhibition. However, with this implementation signals are spatially abstracted and inhibit all other signals, whereas in the classic LIF model a neuron inhibits only its adjacent neurons.

Despite showing high levels of temporal accuracy, leaky competing accumulator models are similar to other accumulation models in that they lack a true spatial component. The LIF algorithm represents a true 2D-map of visual space, whereas accumulator models abstract space into a number of key locations without consideration of their actual proximity. The salience map may be used to influence the parameter choice in these accumulator models (MacInnes, Gorina, Asvarisch, Comardin & Malevich, 2018; Yamaguchi, Valji, & Wolohan, 2018) but the accumulator does not model retinal space per se. Accumulators allows to model experimental results obtained under the laboratory conditions perfectly (Purcell, Schall, Logan, & Palmeri, 2012) but it is not applicable to natural scene processing without first abstracting key locations.

The classic LIF approach is a biologically plausible model that predicts both spatial and temporal aspects of modelling fixations and saccadic eye movements. However, taken with its with default parameters, it has been shown to be limited in simulating temporal behavioural accuracy. Our best results did find a parameter fit that was not different from human data in terms of mean and standard deviation, and to a lesser degree, the overall distributions. The latter however, was only achieved by using the variance between images to generate a realistic distribution, and is a crucial test if we want to evaluate the validity of the assumptions taken by different models (e.g., see Wolfe, Palmer & Horowitz, 2010). Other demonstrations (Ludwig, Farrell, Ellis & Gilchrist, 2009; MacInnes, 2017) have shown that the difference between a bias shift and true improvement in an attentional task can not be seen in mean reaction times, but only in the change in distribution shape.

To sum up, no current model of visual search generates an accurate model of the full response time distributions and spatial locations of saccades. The classic LIF+WTA model produces both spatial and temporal data, but it does not output a temporal distribution of reaction times as it does not have a source of randomness needed to output a distribution. A possible way to overcome this and improve accuracy and comprehensiveness of the model would be the combination of saliency (as a foundation for spatial predictions) and diffusion (for temporal prediction) models for better results.

# References


Badel, L., Lefort, S., Berger, T. K., Petersen, C. C. H., Gerstner, W., & Richardson, M. J. E. (2008a). Extracting non-linear integrate-and-fire models from experimental data using dynamic I–V curves. *Biological Cybernetics, 99*(4–5), 361–370. doi: 10.1007/s00422-008-0259-4

Badel, L., Lefort, S., Brette, R., Petersen, C. C. H., Gerstner, W., & Richardson, M. J. E. (2008b). Dynamic I-V Curves Are Reliable Predictors of Naturalistic Pyramidal-Neuron Voltage Traces. Journal of Neurophysiology, 99(2), 656–666. doi: 10.1152/jn.01107.2007

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). MIT Saliency Benchmark. http://saliency.mit.edu/results_mit300.html

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*.

Carrasco, M. (2011). Visual Attention: The Past 25 Years. *Vision Research, 51*, 1484–1525. doi:10.1016/j.visres.2011.04.012

Davis, L. (1991). *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold.

Gordienko, E., & Macinnes, W. J. (2016, August). Integrated computational model of salience and semantic similarity on spatial attention. In PERCEPTION (Vol. 45, pp. 50-50).

Itti, L., & Borji, A. (2013). Computational models: Bottom-up and top-down aspects. In A.C. Nobre & S. Kastner (Eds.). *The Oxford Handbook of Attention*, pp. 1-20. Oxford: OUP.

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

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. doi: 10.1109/34.730558

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*, 138–146. doi: 10.1016/S1364-6613(00)01452-2

Koch, C. (2004). *Biophysics of computation: information processing in single neurons*. New York, NY: Oxford University Press, Inc.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology, 4*(4), 219–227. doi: 10.1007/978-94-009-3833-5_5

Krasovskaya, S., Zhulikov, G., Merzon, L. & MacInnes, W.J (2018). Training restricted Boltzmann machines to generate human-like eye movements. Poster session presented at the European Conference on Visual Perception 2018.

Ludwig, C. J., Farrell, S., Ellis, L. A., & Gilchrist, I. D. (2009). The mechanism underlying

inhibition of saccadic return. Cognitive Psychology, 59(2), 180–202.

MacInnes, W. J. (2017). Multiple Diffusion Models to Compare Saccadic and Manual Responses for Inhibition of Return. *Neural Computation, 29*(3), 804–824. doi:10.1162/NECO_a_00904

MacInnes J., Gorina E., Asvarisch A., Comardin A., Malevich T. (2018) Capturing attentional capture with salience models. Poster session presented at the European Conference on Visual Perception 2018.

MacInnes, W. J., Hunt, A. R., Hilchey, M. D., & Klein, R. M. (2014). Driving forces in free visual search: An ethology. Attention, Perception, & Psychophysics, 76(2), 280–295. doi: 10.3758/s13414-013-0608-9

MathWorks, (2018). Optimization Toolbox: User's Guide (R2018a). https://www.mathworks.com/help/optim/ug/fminsearch.html

Nelder, J. A., Mead R. (1965). "A simplex method for function minimization". Computer Journal. 7: 308–313.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, *42*(1), 107–123. doi: 10.1016/S0042-6989(01)00250-4

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. Neural computation, 20(4), 873-922.

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

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.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, *7*(1–3), 17–42. doi: 10.1080/135062800394667

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, *77*(1), 157–173. doi: 10.1007/s11263-007-0090-8

Treisman, A. M., & Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Psychology, 12*(1), 97–136. doi: 10.1016/0010-0285(80)90005-5

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. Psychological review, 108(3), 550.

Van Zandt, T. (2002). Analysis of response time distributions. Stevens' handbook of experimental psychology.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural networks*, *19*(9), 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. Vision research, 50(14), 1304-1311.

Yamaguchi, M., Valji, A., & Wolohan, F. D. (2018). Top-down contributions to attention shifting and disengagement: A template model of visual attention. Journal of Experimental Psychology: General.

**Liya Merzon (corresponding author)** lamerzon@edu.hse.ru

PhD student at School of Psychology, National Research University Higher School of Economics.

Research assistant at Vision Modelling Laboratory at Faculty of Social Science and at Center for Cognition & Decision Making, National Research University Higher School of Economics

**Georgiy Zhulikov** geole@mail.ru

Research assistant at Vision Modelling Laboratory at Faculty of Social Science, National Research University Higher School of Economics

**Tatiana Malevich** t.v.malevich@gmail.com

Junior researcher at Vision Modelling Laboratory at Faculty of Social Science, National Research University Higher School of Economics

Werner Reichardt Centre for Integrative Neuroscience, University of Tuebingen, Tuebingen, Germany

**Sofia Krasovskaya** krasov.sofia@gmail.com

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

**Joseph MacInnes** jmacinnes@hse.ru

PhD, Assistant Professor at School of Psychology, National Research University Higher School of Economics.

Head of Vision Modelling Laboratory at Faculty of Social Science, National Research University Higher School of Economics

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

© Merzon, Zhulikov, Malevich, Krasofskaya, MacInnes, 2018