Response to reviewers

We are grateful to the reviewers for their encouraging remarks and useful comments which were essential in improving our manuscript. In addition to the revised manuscript, we detail here our response to the different points being made while showing some screenshots of the revised manuscript when useful. A “revised tracked changes” is also available as a supplementary file to see for all changes being made since the initial revision.

Reviewer #1:

This paper proposes a foveated visual search model. The model implements the What vs. Where separation in a focal accuracy seeking policy (i.e. accuracy driven action selection). The model is experimentally evaluated on a search task for handwritten digits on cluttered backgrounds. Model performance is evaluated and analyzed depending on SNR, eccentricity of the target and the number of saccades performed. Performance as a function hyper parameters is also analysed. The authors made their source code for re-producing the results publicly available. Overall, I think this would be a good contribution to the JOV.

*We thank the reviewer for these positive comments.*

I recommend the acceptance of the paper after the following minor issues are addressed:

- The discussion on the time efficiency of the proposed model over the exhaustive scan (i.e. classical computer vision) approach can be made more convincing. First of all, I think it deserves to be discussed in the main text instead of a footnote (as in page 10). Here is my thinking: suppose that the cost of foveal processing is C. Then, the cost of the exhaustive scan model would be n times C, where n is the number of all pixels (or rather all the locations where fovea will be evaluated). On the other hand, the cost of the proposed model is f times C + f times P, where f is the number of fixations and P is the cost of the log-polar processing model (the Where pathway). The relation between P and n is obviously P = k logn (where k is some constant). Assuming that f is typically much smaller than n, the proposed model seems to be more time efficient than exhaustive search.

*TODO: We thank the reviewer to point out this weakness of our manuscript and have made a consistent effort to shorten, simplify and reduce redundancy in the prose at several places. The most important changes are listed here and all changes are highlighted in the tracked changes PDF. Main points are:*

Minor and more specific comments

- What does "Full-scale" mean in L502? Do you mean high-and-uniform resolution? Please clarify.

*TODO Indeed, this sounds better and as short and we have replaced occurrences of the acronym when possible.*   
  
- In Fig 4, the fonts and graphics are blurry. Perhaps, it is the result of using a raster format instead of vector graphics. Same with Figs 5 and 6.

*TODO Done, thanks.*

- Please consider providing a label for the y-axis in Fig 4.

*TODO: We have improved the overall readability and had the manuscript intensively proofred. Please see the tracked changes’ PDF that highlights all the changes we have done on the manuscript.*

- In the text, figures are cited as "Figure", "figure" or "fig". Please be consistent and follow the journal's style.

*TODO: Done, thanks.*

Reviewer #2:

This paper presents an interesting study at the intersection of neural networks and human vision, specifically involving visual search.

This work is hard to categorize as it sits somewhere between the two aforementioned fields, making direct comparisons to either studies in visual search involving CNNs, or studies involving human participants a challenge. To address this, a suitable paradigm is introduced which involves presentation of the classic MNIST digits at different degrees of contrast over noise. This allows assessment of the degree to which a model that foveates regions of the image (modeled by a log-polar transform) can effectively localize and identify targets of interest.

For me the key finding of this work is that a sub-linear optimized spatial search is useful and effective in localizing and identifying targets of the type chosen.

*TODO: We thanks reviewer #2 for his encouraging and valuable comments. In our revision, we have tried to put forward these strong points and render them more visible to the readers. Let us only precise that we prefer to refer to our two experimental measures (anticipatory pursuit and confidence ratings) as implicit and explicit respectively, rather that unconscious and conscious, as we do not have a rigorous assessment of the unconscious nature of anticipatory pursuit. This point is also the object of an observation of Reviewer 1*.

I do have some suggestions and questions relating to the manuscript as follows:

i. The overall presentation could be tightened up a little in terms of grammar and sentence structure

*TODO: Indeed, Figure 3 synthesizes in one figure all traces: the probability-bias implemented following the generative model, individual ratings, anticipatory eye velocity and Bayesian model predictions. In the revised version, while keeping the overall architecture for the figure, we have tried to make it more readable and tried to improve its caption. Instead of showing the stacking of all modalities for two subjects (with one panel per subject), we now show one panel per modality (probability-bias / anticipatory pursuit / rating) for the whole group of n=12 participants. This better illustrates the methodology in the paper which confronts experimental data with theoretical predictions:*

ii. Figure 4 references orange bars, but they appear to be brown to me.

*TODO: Similarly, Figure 5 was cluttered with too much information (see also the comment from reviewer #1). We have simplified it by showing the analysis (estimated hazard rates) for each individual participant. Last, we have added a panel showing the response of the model for three characteristic levels of the hazard rate and corresponding to the range of volatilities observed across participants.*

iii. The notion of optimal strategies for exploration could be expanded upon. E.g. The discussion of Najemnik and Geisler's work is a good fit, one could also include Bayesian Surprise (The IK reference is inappropriate - this was Baldi and Itti), or other information seeking strategies - the AIM or SUN models)

*TODO: We are grateful to the reviewer to have raised our attention to several points lacking clarity. We have now more precisely described the experimental trial design, and explicitly mentioned the step-ramp paradigm at lines 891-894. As the initial target position is randomized across trials and not disclosed until the end of the gap, we can exclude that it is used as a cue for target motion.*

iv. Central to the model is the decision between foveation and identification. The paper states: "If the predicted accuracy in the output of the "Where" network is higher than that predicted in the "What" network, the position of maximal activity in the "Where" pathway serves to generate a saccade which shifts the center of gaze. Else, we interrupt the visual search and classify the foveal image using the "What" pathway such as to give the answer (ANS)." It is not clear that these quantities are on the same scale or comparable. More detail on this particular mechanism would be welcomed.

*TODO: We appreciate this suggestion by the reviewer and we have now added a small paragraph (lines ...)*

v. I wasn't able to discern whether the network is trained piecewise (e.g. with the saccade decision part done manually in code), or whether the entire network is trained with the BCE and argmax end-to-end. If slightly more detail could be provided on the nature of the training procedure and how one constructs the model in such a way that it performs both selection and identification while training, this would also be welcomed.

*TODO: Done, thanks. Please see the tracked changes’ PDF that highlights all the changes we have done on the manuscript.*