Masoud Ghodrati,
Neuroscience Program, Biomedicine Discovery Institute,
Monash University, Clayton, VIC., Australia
E-mail: `ghodrati.masoud@gmail.com`
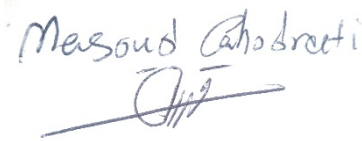
Journal of Vision
USA
The editors

Melbourne, 25/5/2019

Dear **Editor**,

The authors would like to thank the respected reviewers and editorial members who kindly participated in the review process. Please find enclosed the new version of our manuscript "Object categorization in visual periphery is modulated by delayed foveal noise". We carefully took into account the reviewers' comments which helped us to increase the quality of the manuscript. All the suggested refinements are performed and the manuscript is revised accordingly. We hope that the new version of the manuscript is suitable for publication in *Journal of Vision*.

Yours sincerely,
Masoud Ghodrati, PhD

Reviewer #1

The authors have met all my expectations, and I'm happy to accept their paper for publication. There is only one small correction I would suggest, but I will leave it to the authors to decide. The authors mention "the well-established finding of parafoveal preview (Henderson et al., 1989; Henderson, 1992), where parafoveal presentation of visually related flankers to the target object facilitates identification of the target." This doesn't sound quite right. The effect is that, if a peripheral target is attended to prior to making a saccade to it, parafoveally processed visual features of that target can facilitate recognition of the target after the eyes land at it. This is shown by the fact that, if the target changes during the saccade to it, if the new target shares visual features with the previewed target, recognition of the new target is facilitated relative to cases in which the new target is visually dissimilar to the previewed target. I don't think it is so important that the paper needs to be revised, and will leave it to the authors to decide.

> **Reply**: Thanks for your precise view on the manuscript. To make it clear we changed the sentence to:

> *"This is consistent with the well-established finding of parafoveal preview (Henderson et al., 1989; Henderson, 1992), where it is shown that by attending to a peripheral target before making a saccade, the parafoveally processed visual features of that target can facilitate recognition of the target after the eyes are landed on it."*

Reviewer #2

I'd like to thank the authors for their attention to my comments, and I appreciate the work they put in to improving the paper; I think it is much clearer in the text and plots. Here, I only address issues that require more follow up. My words are in green, and theirs in black.

"Reply: We corrected the citation to Rosenholtz 2016. Moreover, we added the crowding to the introduction (page 2, line 36):"

The authors improved the discussion of crowding significantly. I am not satisfied, however, with aspects of how the authors implemented it. First, they should focus on how crowding is relevant to the results. I should have made this clearer in my review. In their rebuttal, the authors state, "here we did not aim to measure the effect of crowding on subjects' performance" which I agree with. I disagree, however, with their next statement as a dismissal of crowding, "It is also important to point out that making a direct relationship between crowding and our results is a bit complicated as we presented the same object images in different categorization tasks – i.e., the same objects could accurately be categorized in superordinate level but subjects were less accurate in categorizing the same images in subordinate level" This argument works equally well if you replace "crowding" with "acuity"; and they do make direct relationships to lower acuity. Ultimately, this study cannot tell apart whether performance is lower in the periphery because of crowding or acuity; that would require more experiments as the authors suggest in lines 282-283, or an analysis that directly compares performance in this experiment with performance in experiments where subjects foveally viewed low-pass stimuli.

> **Reply**: We agree with the reviewer that extra experiments directly comparing the effects of acuty and crowding on object recognition are required and the performed experiments cannot tell how each of these two factors impair the fine object categorization in periphery.

I strongly suggest they replace this list of facts about crowding with a shorter statement, appended to the previous paragraph. For example, something like this would be adequate for the introduction:

"Moreover, it has been shown that the task demand (e.g., object detection or identification) modulates the object processing in the processing of objects in peripheral vision (Jebara, Pins, Despretz, & Boucart, 2009). It is important to consider the potential role of crowding in categorization as well. Behavioral studies in humans show that although fine details can be recognized in isolated peripherally presented objects and scenes, perception is impaired when those stimuli are surrounded by clutter or internally complex. While the mechanism that leads to it is still an active topic of research, crowding is the most important limitation in peripheral vision (Rosenholtz, 2016)."

If the authors insist on keeping the list of models of crowding (which I think is not relevant), they should also include Rosenholtz's forced texture perception alternative (e.g. Balas et al., 2009; Keshvari & Rosenholtz, 2016).

Reply: As truely pointed by the reviewer, we eliminated the unnecassary facts about crowding in introduction. To do so, we replaced this paragraph with the reviewer's suggestion.

In the Discussion. Lines 265-269 ("The lower… disrupt the recognition accuracy"), it is very unclear whether the authors are invoking acuity or crowding to explain the results. I would recommend something simpler, such as the following:
"The performance drop in the periphery is likely due to an increase in crowding and decrease in spatial acuity, which respectively jumble and reduce the spatial details necessary for categorization. It is important to note that 'self-crowding' can occur with a single object (Martelli, Majaj, & Pelli 2005), which may be at play with our natural image stimuli."

Reply: We replaced these sentences with the reviewer's suggestion. Now it clearly explains that low acuity and crowding can both be the causes of accuracy drop in visual periphery.

I think the discussion about Wijntjes & Rosenholtz is not applicable here for the following reason. The current study takes care to select images in which the context is minimized, by first selecting images that have a large object, and then cropping them around the central object. In Wijntjes & Roseholtz, they test contexts that take up to most of the visual hemifield, and they find only a small increase in performance for the smallest contexts; even the smallest contexts in their paper are much bigger than the contexts in the stimuli in the current paper. For this reason, they cannot use Wijntjes & Rosenholtz to weaken the potential role of crowding.

Reply: We think it is important to cite and discuss Wijntjes & Rosenholtz's paper in our discussion (as suggested by the reviewer). To avoid misleading information, we changed the conclusion as below:
"*Although in our experiments, all objects were presented in their real-world congruent background and we cropped the images in a way that the target object covers a large portion of the image, crowding still remains one of the possible reasons for the accuracy drop in periphery.*"

The paragraph in Lines 275-283 seems to be a repeat of the previous paragraph; I suggest only keeping the last two sentences of it (Lines 280-283) and appending it to the previous paragraph.

Reply: As suggested by the reviewer, we kept the last two sentences of this paragraph and eliminated the remaining.

"Reply: How did the authors choose their stimuli from the huge dataset? Was it randomly, or was there some selection process?"
Line 113-115: I appreciate that the authors made more specific their selection of the stimuli. For this study to be replicable, though, the authors cannot just say that they "selected a larger set of images in which the target object covered a large portion of the image, we then randomly selected the final image set from the initially selected set"; they must give a deterministic way to come to same images. I suggest either releasing the subset of images they used online, or the minimum percentage of image covered by the object necessary to be included in the study. For instance, I would not be able to select which images had a large enough object given this imprecise description.

Reply: The first selection phase (i.e., selecting images in which objects covering a large portion of the image) was to prevent the object scale variation affecting the recognition performance,

actually we did not want to impose another term of difficulty apart from the categorization level. As suggested by the reviewer, we will upload all of images on github available at `https://github.com/Masoud-Ghodrati/vis_peri` after acceptance. It is mentioned in the manuscript as well.

Similarly, lines 152-153: "(i.e. in some locations 44 images…46)" is unnecessarily imprecise; please replace with a range or standard deviation. Is it 40-50? 30-60?

Reply: The range was from 44 to 46. We added this range to the text.

In lines 309-317, the authors use the fact that the 300 ms delayed noise caused the largest impairment in basic and subordinate categorization, along with the statement that some saccades happen 300 ms after stimulus onset, to imply that the two have something to do with each other. This is problematic for two reasons. First, just because these two things happen at the same time scale does not imply that "this accuracy drop could probably be due to inconsistency"; there has to be a more causal explanation, which might have been tested by other studies. Second, while I was unable to access Fischer 1987, Thorpe et al. 1996 only show that median "go" reaction times for their task range from 382 ms to 567 ms, while the brain signal that distinguishes go vs no-go trials occurs 150 ms after stimulus onset; there is no particular coincidence with the current paper's 300 ms finding, nor does the Thorpe paper mention saccades or use peripheral stimuli. I think this logic must be better justified or removed.

Reply: You can find the Fischer 1987 paper in the link below:
`https://drive.google.com/file/d/1E6175qQExYj7IVWlZlyy82A6FqtCZ10x/view?usp=sharing`
Fischer 1987 states that a regular saccade in humans is around 200 ms (150-250) and they show that in some special conditions it can be even less. Thorpe et al. 1996 shows that the EEG signal of subjects doing animal/non-animal (i.e., superordinate level) task is distinguishable around 150 ms. By putting these together one might say that processing the target object in periphery and then making a saccade toward it may need around 300 ms. To avoid making a strong claim, we updated the sentence as:
"*While this result should be interpreted with caution, it nevertheless suggests that this time window (300 ms) may be important for resolving peripheral representations, perhaps prior to a goal-directed saccade that would eventually foveate the target object (150 ms for the initial object processing (S. Thorpe, Fize, & Marlot,1996) and 150-200 ms for making the saccade (Fischer,1987))* "

Finally, I noticed several grammar and spelling issues, I still recommend the authors have a colleague or third party read the paper for mistakes. They are listed below.
The first line of the Abstract "Despite the lower…" still reads awkwardly, and "acuity" should really be "fidelity" given that crowding is the main loss in peripheral vision.

Reply: We asked our colleague for proofreading and editing of the manuscript. We changed this sentence to:
"*Behavioral studies in humans indicate that peripheral vision can do object recognition to some extent.*"

Title: Would sound better without the "a" before "delayed"

Line 5-6: "retinotopic to fovea" should be "retinotopic to the fovea"

Line 7: The word "humans" should be removed or replaced with "human"

Line 10: "foveal representation" should be "the foveal representation" or "foveal representations"

Line 20: "The foveal" should be "Foveal"

Line 32: Replacing the word "capability" with "capacity" would improve the sentence

Line 38: The first "The" should be deleted

Line 82: "periphery, so, the information…" should be "periphery, such that the information…"

Line 81: The sentence beginning with "Here, we hypothesize…" has awkward grammar; I suggest replacing "…available to peripheral vision…" with "…available to peripheral vision is sufficient to support a general guess about an object's high-level category (animal vs. nonanimal)"

Line 90: "involve" should be "involves"

Line 107: ", which was forced by foveal noise." sounds awkward, I suggest "using foveal noise."

Legend above Fig. 3C: "fighterb" should be "fighter"

The gray bars in figure 5 are a bit difficult to see on the bars, perhaps black would be more visible?

I think it is a bit misleading to connect the "No noise" data points in Figures 4 & 5 to the "100 SOA" data points with lines; this still implies there is a continuum between no noise and noise, which is not true.

Line 160: "(continuously changing)" should be moved to the first time they mention the dynamic noise mask in line 156.

Line 188: "the central vision" should be "central vision"

Line 230: "can be due…" should be "may be due…" or "might be due…"

Line 252: "at the very far periphery" sounds better as "in the very far periphery"

Line 279: "periphry" is misspelled

Line 280: same typo as previous line

Line 287: "Therefore," is a bit awkward, may "To this end," is better

Line 310: "making" should be "make"

Line 331: "form" should be "from"

Lines 331-333: I suggest "first, second, and third" instead of "firstly, secondly, and thirdly"

Line 337: "and what is the temporal dynamic" would sound better as "and the temporal dynamics"

Reply: As the reviewer suggested we corrected all the grammatical mistakes and updated the images needed to be changed.