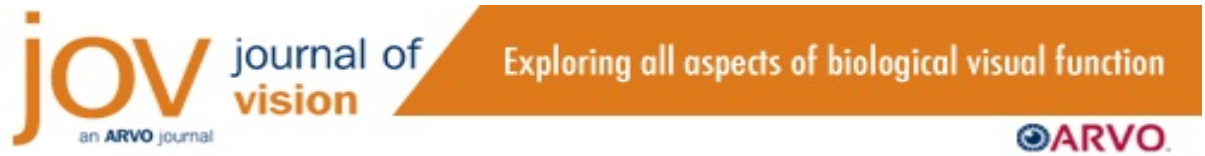
Subject: Journal of Vision - Decision on Manuscript ID JOV-07646-2020R1
Date: Thursday, April 15, 2021 at 7:04:47 AM Central Daylight Time
From: jov@msubmit.net
To: zachary@neurophysicole.com
CC: felix.wichmann@uni-tuebingen.de, Karl Kuntzelman, Michael Dodd, Matthew Johnson, Zachary Cole



Manuscript ID JOV-07646-2020R1 titled "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements"

Dear Mr. Cole,

Your manuscript has been carefully reviewed and is potentially acceptable for publication, but there is still a conceptual issue in the discussion---the relevance of image content---raised by reviewer #2 that needs to be addressed before the paper can be accepted in its final form. The reviewer comments are appended below. Please prepare a revised manuscript and a point-by-point response to each issue identified by the reviewers. Only after satisfactory revision will your paper officially be accepted for publication. Please try and return your revised manuscript within 30 DAYS.

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text.

Please prepare a point-by-point response to the suggestions of the reviewers. This can be a Word or PDF file to be uploaded with the rest of the manuscript files under the "Author Response to Reviewer(s)" file type. Please be as specific as possible in explaining the changes made to your manuscript.
In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s).

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

If it is not possible for you to submit your revision within the time requested please contact the journal as soon as possible.

Once again, thank you for submitting your manuscript to the Journal of Vision and I look forward to receiving your revision.

Sincerely,
Felix Wichmann
Editorial Board Member
JOV

Editor, Journal of Vision

****

Editor Comments:

Reviewer #1 (Comments for the Author (Required)):

I am not fully convinced by the authors' rebuttal against my comments on the Yarbus problem, however, I thank the authors for explaining and comparing their approach in revised manuscript appropriately. Thus, I am satisfied with the revision.

Reviewer #2 (Comments for the Author (Required)):

I want to thank the authors for their extensive reply to my review. I don't see any remaining problems with methods and results, but I think some conceptual points still need more discussion and justification. Below, I'll first go through their answers to my concerns one by one, and then also list some other things that I noticed while reading the revised manuscript.

The notions of "tasks", "cognitive processes" and "mental states"

I thank the authors for the clarification of the terms used, however, the definitions still seem to be a bit vague to me, especially the difference between cognitive processes and mental states. I would probably define the terms slightly differently: For me, a process includes a dynamic (it processes something) while a mental state as such is static (after all, it's called a state...). This intuition aligns with the notion of dynamic systems in mathematics, where a dynamic process evolves the state of a system over time, while the state describes where the system is at a given point in time. In the case of a mental state, the state includes which processes are ongoing and additional information. I would probably also disagree with the definition of cognitive processes as a theoretical construct. To me, they seem to be quite real, although hard to measure: When trying to answer a question about an image, I can observe myself going through such a cognitive process (e.g., for the question about the wealth of the people in the image).

Whether using my definition or the author's definitions, I think it's best to avoid talking about classifying mental state from eye movement data. While this might be possible in theory, most likely it's infeasible in practice (as the authors point out, it would involve mood, present memories and many other things). But we can clearly classify tasks from eye movement data and, at least for my definition of cognitive processes, it might be possible to understand the cognitive process at work when solving a task.

The relevance of image content

I thank the authors for their extensive response to my concern. Unfortunately, I think I still consider the stimulus quite relevant for differentiating task from eye movements. First, since maybe this was not clear enough from my first review: I want to emphasize that when talking about the relevance of stimulus information for eye movements, I'm not referring only to low-level visual information (e.g., in the sense of the classic saliency since Treisman&Galade and Itti&Koch). I'm referring to image content in general, including objects, semantics and everything that observers are able to recognize in the image. This especially includes the case where different tasks are used on the same image (see l.119 of the revised manuscript). I have a hard time imaging that the main reason for differences in e.g. saccade length distributions between the task of memoization and the task of image rating is anything else than a difference in the relevance of different objects/regions in the image, their spatial arrangement, and the best order of information acquisition. For assuming that the image is not relevant, one would have to assume that people actually move their eyes not in a way that looks for information about the image, but in a way that is independent of the image (but dependent on the task) and just accumulate the information that the eye happens to attend. Since, e.g.,

relevance of different objects changes across tasks, this makes the relevance of image content I'm worries about not a bottom-up effect, but a top-down effect.

I think there is an additional argument that image content can and maybe even should part of the inverse Yarbus problem: Yarbus himself never only shows eye traces. All his figures also show the observed image, especially the famous Figure 109 of the unexpected visitor. Here, most likely, I would be possible to differentiate at least some of the tasks in this figure just from the gaze traces (mainly by checking the overall scattering of fixations), but it gets much easier and at the same time much more interpretable if I know the image and notice that, e.g., in panel 3 (age of people) the observer focuses on the heads, while in panel 5 (clothes of people), the observer focuses more on the bodies. Whenever I think about why the gaze traces in this figure look like they look for a given task, I think "of course the observer looks at these image areas, because they contain the information that is needed to answer the question". It's easy to imagine tasks that would be completely impossible to differentiate purely from gaze data: We just have to make sure that there are two different sorts of objects in the images which follow an identical spatial distribution and have one task be about the first kind of object and the second task about the other kind of object. For example, to follow up on Yarbus' tasks, we could ask observers either about the average age of women or the average age of men in the image. Without knowing the observed image (and therefore, whether a given observer looked more at women or men), it should be close to impossible to differentiate these tasks.

So, in the end I think that for truly understanding how observers approach any of the discussed tasks (i.e., why they behave the way they do and which cognitive processes are ongoing), there is no way around taking into account what's in the image. However, I am aware that this would be a major extension of the presented study, which would need substantial additional work for getting access to the relevant image information and therefore might be very hard or even infeasible. I still see value in the present study, i.e., in better understanding how well tasks can be decoded purely from eye movement data without image information. I just think we cannot expect to learn too much about to underlying cognitive processes (except via, e.g., post-hoc analyses of the data). Therefore, I'm fine with leaving a more complete approach combining both eye movement data and image information for future research. But I think even in this case it should be acknowledged in more detail that and how image content will affect gaze traces including in the datasets at hand (unless the authors convince me otherwise, of course), and some reasoning why it is okay to not take this effect into account in the present study (e.g. something along the lines of what I wrote above).

comparison of timeline and image model

I thank the authors for the clarifications. I was not aware that the image model and the timeline model are quite comparable in terms of parameter count. This makes the results indeed more interesting. If the question is which data format makes the task information easier decodable, it might be interesting to add a third model which as access to both timeline and image data. After all, it is possible that the information extracted by the timeline model and the image model is not identical. In this case, it might be possible to achieve better information in a joint model. If the joint model performs as well as the better model (i.e., the timeline model), this indicates that the image model has not access to any helpful information that's not already accessible to the timeline model. However, this is only a suggestion. I don't think it's necessary for the paper to be interesting.

Comparison to other approaches

I appreciate the authors' effort to compare to other methods. It's interesting to see that one of the best-performing previous methods didn't work at all on the present dataset.

Other points of first review

I'm happy with the answers to the other points in my first review and I won't go into detail about them

Additional points

- I noticed I'm not really sure how many images the two datsets contain. I guess each trial corresponds on one image,

therefore the exploratory dataset contained a total of 120 different images?

- In lines 403-407 the authors compare the classification accuracy on the exploratory and confirmatory dataset and don't find a significant difference. I just want to point out that even if they did find a significant difference, this would not necessarily indicate a generalization problem. The confirmatory dataset uses a slightly different search tasks which now includes a cue. This will most likely affect the gaze traces, which should now focus on fewer image areas. In turn, this could make the task classification itself both easier or harder, resulting in a different achievable accuracy. Therefore, in my opinion, this statistical test could be removed. The most important insight of the confirmatory dataset is that all relevant qualitative effects where reproduced, which seems to be the case.

- I want to applaud the authors for making data and code publicly available.