UNIVERSITY OF Nebraska Lincoln

DEPARTMENT OF PSYCHOLOGY

15 May 2021

Editorial Office, *Journal of Vision*

Dear Dr. Wichmann and colleagues,

We would like to start by again thanking the editor and reviewers who took the time to carefully review and evaluate our manuscript "Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements" originally submitted to *Journal of Vision* on September 14, 2020, then revised and re-submitted on February 17, 2021. We were encouraged by the enthusiasm for our work that was expressed, and we have now prepared a revised manuscript for your further consideration. Overall, we feel that the changes we have implemented in response to the thoughtful comments and recommendations from the editor and reviewers have improved the quality of the manuscript.

The enclosed document provides a verbatim copy of the initial review comments (written in black text) and our responses recorded in-line (in blue text). We made the effort to respond completely and concisely to each comment, in addition to providing a short description of how the comment or recommendation is addressed in the revised manuscript. In the revision, all changes are shown in red text.

Once again, we are grateful to the editor and reviewers for their time and consideration. We hope that these changes adequately address all issues raised by the reviewers and that you will now find the paper suitable for publication.

Best regards,
Zachary J. Cole and colleagues

238 Burnett Hall / P.O. Box 880308 / Lincoln, NE 68588-0308 / (402) 472-3721 / FAX (402) 472-4637

Reviewer #1

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.

We thank the reviewer for taking the time to review our revised manuscript, and appreciate their acknowledgements of our efforts to review the manuscript despite still disagreeing on a theoretical point regarding the nature of the Yarbus problem.

Reviewer #2

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.

We appreciate the clarification by the reviewer. We agree that referring to the classification of task rather than mental state is more accurate and specific. We have adjusted the manuscript so that it mentions the classification of task rather than mental state.

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).

Again, we would like to thank the reviewer for taking the time to clarify a point made in the original review. We tend to agree with the reviewer that the content of the image can be important to understand and interpret gaze data with respect to top-down influences on attention. In this particular study, we did not take into account the image data because the images were randomly assigned to conditions between participants in both experiments, a method that should at the very least reduce any bias in the eye movement data that was introduced by objects or semantic information unique to any individual image. If we were to tie the tasks to a particular image type, the nuances of the top-down mechanisms at play actually have the potential to be washed out by the lower-level bottom-up factors associated with the stimulus, as pointed out by Borji and Itti (2014). All of that said, we agree with the reviewer that including the visual information from the image

would be a logical and potentially informative next step, which we acknowledged in the original version of the manuscript (see lines 534-542 of the previously revised manuscript).

With regard to line 119, which was mentioned specifically by the reviewer we were meaning to refer to a portion of the literature that has classified task using paradigms that require the participant to read a paragraph or view an image. These tasks result in easily distinguishable eye movements which can be decoded and classified, but the mechanisms differentiating these eye movements are, in our eyes, more attributable to the structure of the image than to the impetus of the participant. For this reason, we believe that although these paradigms have provided a valuable addition to the literature, they do not fit what we believe to be the inverse Yarbus problem. The reviewer's comment has pointed out to us that the language used in the concise clarification that we added to our previous revision (line 116-120 of the previous revision) needed to be more specific, particularly with regard to the concerns raised by the reviewer. For this reason, we have revised our previous revision to more accurately reflect what we believe to be the sentiment of the reviewer.

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.

We appreciate this suggestion from the reviewer. While we agree that the suggestion of a third model would be intriguing, currently, the two approaches result in separate data formats that could be more-or-less suited to the decoding algorithm. For this reason, this approach might necessarily require its own line of investigation (a completely new project). As such, we have not updated the current manuscript with this information, but we fully agree that this would be an interesting avenue for investigation, and have added this to the potential queue of analysis to follow-up this current project.

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.

We would like to thank the reviewer for the acknowledgement of our efforts. Although the process was not simple or straightforward, we gave our best effort in attempting to find another study with code and data that we could adapt for a direct comparison with our own approach. We also found it interesting that our approach outperformed the Coco and Keller (2014) approach. We believe this outcome can be partially attributed to the focus of Coco and Keller being on distinguishing the task sets based on the expected underlying operations that take place rather than relying on distinct features in the data or a complex model architecture. With the expectation that future readers will also be intrigued by the outcome of this replication analysis, we took the liberty of adding this caveat to the manuscript (see lines 457-460).

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?

Yes, each trial corresponds to an image. Each participant viewed 120 images; but these images were not all presented in the same order, and were not all corresponding to the same tasks between participants. Overall, the dataset only contained 120 images, but each image was randomly assigned to one of the three conditions across all of the participants in each experiment.

- 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.

This is a valid point made by the reviewer. We agree that if the classification accuracies for these two experiments was different, that might not necessarily be an indicator of a problem with generalization. Based on these comments from the reviewer, we agree that the overall replication across the two studies is the more important insight, but we do believe that analysis in question does show some continuity between the two experiments that can be an indicator of a stable approach. For these reasons, we did not remove the analysis, but removed all language suggesting that the outcome of the analysis comparing classification accuracy between the exploratory and confirmatory datasets was an indicator of generalizability.

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

Thank you! Based on the difficulties we experience attempting to replicate other approaches, we believed this making the data and code publicly available would be of greatest benefit to the field as a whole.