Project Proposal
Prediction of imagined and perceived complex visual stimuli in a VR environment based on eye movements

Ana Dobre

BSc (Hons) Computer Science
The University of Bath
November 2020

Contents

Chapter 1

Introduction

1.1 Problem Description

This project aims to improve on performance of a recent *Nature* publication (Wang et al., 2020), which attempted to computationally discriminate a visual image amongst 100 images based on gaze patterns from observers perceiving those images, and apply these classifiers to gaze data from recall/visual imagery of the same stimuli. Whilst Wang et al (2020) achieved impressive results for discriminating perceived images using K-Nearest Neighbour and CNN classifiers, performance for visual imagery was poor. Low performance is likely driven by several factors such as; scaled, shifted and translated gaze areas during imagery compared to perception; data was reduced into histograms, however two distinct images can result in very similar histogram, reducing discriminability; participants had their eyes open which results in interfering perceptual information during the imagery task.

This project aims to conduct a similar study, however in a virtual reality environment to improve performance. Doing so will reduce interfering perceptual input, participants can keep eyes open to enable eye-tracking, but the environment be made pitch-black. Further, the large-scale environment in VR, could be exploited by using upscaled cues, to prevent the issues of downscaling eye-movements in the imagery conditions reported by Wang et al. Moreover, in a 3D environment, depth components of visual imagery can be explored, for example whether eye movements are similar in depth between perception and imagery.

1.2 Background

Relation between imagery and perception

Eye movements are usually directed to the same stimuli that catch one's attention (Holmqvist et al., 2011). Over time, a strong correlation has been shown between

voluntary eye movements during perception and involuntary eye movements during imagery of the same images. Brand and Stark (1997) used images with minimal complexity (irregularly-checked grids) to demonstrate the similarities between eye movements during memory retrieval and encoding. They concluded that those eye movements reflect the content and spatial layout of an imagined scene. They also appreciate that mental imagery uses mechanisms similar to perception and that the involuntary eye movements have a role in recalling separate parts of a complex scene and arrange them in a layout that resembles the original whole scene, this is a theory supported by Hebb (1968), Laeng and Teodorescu (2002) and Mast and Kosslyn (2002). In contrast to this interpretation is Pylyshyn, 2002, 2003 where it is argued that there are no internal images and that human mental representations are propositional. He also claims that the mental representations of objects depend on other spatial indices from the environment.

Johansson et al., 2006 propose an experiment that contains two different types of environment: one light and the other completely dark. The study concluded that imagery could not be strictly linked to spatial indices because participants performed similarly in both environments. Another conclusion from Johansson et al., 2006 is that participants made involuntary eye movements recalling verbal and visual stimuli, and the eye-movement effect was equally strong for both verbal and visual stimuli. This effect was tested using complex stimuli, but the variation was laking. The study mentions that if excessive blinking is removed (to limit gaps in gaze data) and a frame of reference is provided (to avoid frequent recentering and resizing), the pitch-black environment can lead to better results compared to when participants were looking at a whiteboard. Creating a pitch-black environment is desirable also because it has been shown that people prefer recalling an image with closed eyes (Vredeveldt et al. 2015, Mastroberardino and Vredeveldt 2014).

Involuntary eye movements and encoding eye movements are similar, but not identical

Involuntary eye movements operate as a functional role in memory retrieval, but they are not reinstatements of those produced during encoding (Richardson et al. 2000, Johansson et al. 2006,2012 and Brand and Stark (1997)). In the same studies, involuntary eye movements during imagery appear scaled and translated compared to the movements made during perception. Johansson et al., 2005 suggest that this phenomenon occurs to allow the participant to view the vast majority of the image as a whole during the recall process. Kozhevnikov et al., 2005 and Johansson et al., 2006 observed that people with scaled and translated imagery eye movements were also the participants who had high scores on spatial imagery. Same studies explain that eye movements during visual imagery tasks are employed to reduce cognitive resources associated with the processing of spatial information, and a weaker spatial imagery ability increases the need for those eye movements.

Johansson et al., 2012 analyse how eye movement impairment activates the recall capacity of a visual image. Participants were asked to recall spatial arrangements while looking at a blank screen, an area that was associated with the image, an area that was

unrelated to the original image, or while fixing a cross. Scholz et al., 2014 conduct a similar study using verbal stimuli associated with a specific area in the environment. In one of the tasks, the participants were asked to direct the gaze towards the relevant area while in the other task, the eye gaze was directed to an area unrelated to the encoding moment. In both cases, the performance was better when participants kept the eye gaze on the same area during the encoding-recall process. Therefore overlapping or maintaining the reference frame between encoding and recall leads to more accurate results. Moreover, both studies suggest that looking at nothing and memory retrieval have a functional relationship and it is also speculated in Scholz et al., 2017 that these eye movements are driven by covert attention.

Image retrieval methods

Kosmyna et al., 2018 developed an EEG-based Brain-Computer Interfaces that distinguishes between visual perception and visual imagery signals, and decides which visual stimulus is used. They also aimed at distinguishing between rest versus imagery and rest versus observation. The classification accuracy leads to poor results (between 61% and 77%) but nevertheless, these results show that visual imagery can broaden the range of BCI control strategies.

A communication-BCI proposed by van den Boom et al., 2019 uses visual imagery to detect letters as a fast and intuitive way of spelling. The decoding of visually imagined characters has an accuracy significantly above chance level. Shen et al., 2019 developed a deep neural network capable of directly mapping fMRI activity to the perceived stimuli and therefore reconstructing the perceived image from fMRI data during perception. The general layout and sometimes colour of the stimuli were preserved in the resulting images. The neural network was trained only on natural images but performed well when letters or abstract shapes were used. However, these approaches involve a large number of motor restrictions on participants.

In comparison with EEG and fMRI, using eye gaze to observe involuntary eye movements during memory retrieval is a more natural approach, which can be embedded in other systems, giving participants more freedom. Wang et al 2020 conducted a study that aimed to discriminate between images based on eye gaze during imagery and perception. This is the only study aiming to retrieve a mental image based on involuntary eye movements. They concluded that this is possible and their study is discussed in the next section.

1.3 Existing solutions

Exploiting the similarities between eye movements during encoding and recall, Wang et al (2020) attempted to computationally discriminate a visual image amongst 100 images based on gaze patterns from observers perceiving those images and apply these classifiers to gaze data from recall/visual imagery of the same stimuli.

Figure 1.3.1. Method used when participants were asked to encode and recall an image (Wang et al., 2020)

Image retrieval requires high similarities between eye moves of different participants while looking at the same image. In this study, 100 naturalistic stimuli were used (a database large and variate enough compared to visual stimuli used in previous studies). Since the degree of similarity between encoding and recall eye movements was not quantified before the following scenarios were used to assess this:

Scenario 1 Computationally discriminate a visual image form 100 other images based on encoding gaze patterns

Scenario 2 Computationally discriminate a visual image form 100 other images based on recalling gaze patterns

Scenario 3 Using a classifier that performs well in the first two scenarios, discriminate against a new set of images using new gaze patterns.

During the experiment, 28 participants were asked to look at a picture for 5 seconds and then recall the same picture while looking at the same screen (now empty) for another 5 seconds in a dark and quiet environment as shown in Figure 1.3.1. 200 gaze patterns (half for recall and half for encoding) were gathered from each participant. The gaze patterns were then transformed in 2D density histograms.

For classification of gaze patterns, k-nearest neighbor was used as a baseline for comparison. The accuracy of this classification method when retrieving an image based on encoding movements was 94.5%, but for recall movements performed poorly (54.3%). Convolutional Neural Networks were also used for image retrieval in the first two scenarios, following the structure from Wang (2019). The accuracy of image retrieval based on encoding using CNN was 97.5% and 69.8% was achieved when using recall data. Due to distortion during recall, they mapped imagery to perception to create a decoder within the network. When an imagery histogram was fetched, the decoder would create a corresponding encoding histogram. The accuracy registered after adding the decoder was 72.1%. For the third scenario to retrieve an image, recall eye movements were matched with encoding eye movements from the same viewer. The encoding sequences were already mapped to the right images. After this, a recall histogram was assigned to the class of the matching encoding histogram with an accuracy of 66.4%.

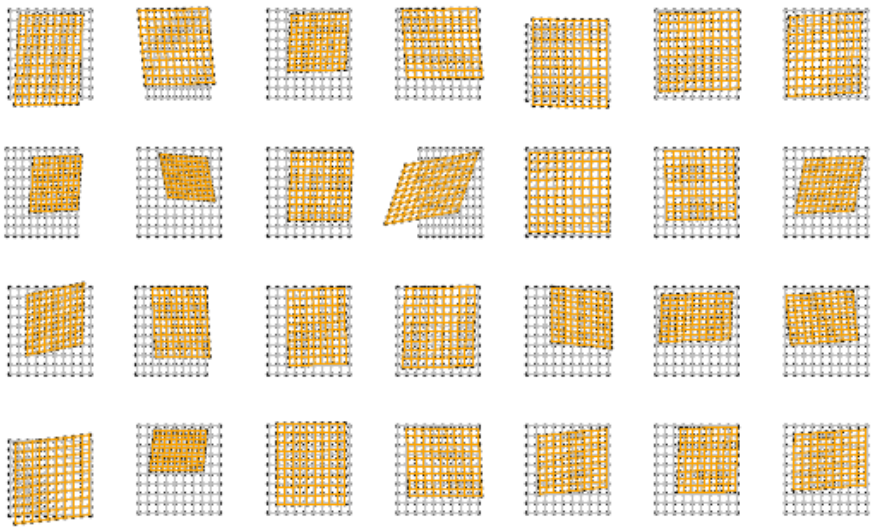

Figure 1.3.2. Example of distorted gaze patterns during recall (yellow) compared to gaze patterns during encoding (gray)(Wang et al., 2020)

Two of the main limitations that lead to poor classification performance were the distortions in recall gaze patterns as shown in Figure 1.3.2 (the discrimination performance also varies between individuals) and the use of histograms which inevitably lead to similar data sets for different pictures. Even though they experimented with different kinds of histograms, the results were not significantly different. Considering the current approach this can lead to difficulties when extending the stimuli database. The distortions in recall gaze data were likely an effect of a missing reference frame. The modest improvement of 2.3% after mapping the imagery to encoding gaze patterns might be explained by the lack of consistent deformation between the recall gaze patterns even in data sets of the same participant.

Wang et al (2020) conclude that it is possible to retrieve an image based on encoding and recall eye movements with very few restrictions imposed on participants (unlike in similar previous fMRI studies).

1.4 Proposed Solution

1.4.1 Aims

Replicating Wang et al., 2020 Nature study in virtual reality with improvements to the prior methodology. Aiming to achieve accurate classification of visually imagined, and visually perceived stimuli, based on eye-movement data and exploration of depth information in

visual imagery compared to perception.

However, considering the health restrictions that might be imposed in the future, the project might need to shift from the virtual reality environment to a "normal" display which would make the exploration of depth information difficult. The setup in this case would contain a laptop or desktop and a desktop-based eye-tracker.

1.4.2 Objectives

1. Implement a VR scene adapted from the study.
2. Make use of gaze-tracking to detect involuntary eye movements.
3. Areas to improve/explore based on the Discussion section from Wang et al., 2020:

- *Remove distracting external perceptual input:* VR can facilitate a pitch-black environment, but frequent blinking must be dealt with to reduce the gaps in the data collected by the gaze tracker.
- Experiment with different classification pipelines.
- *Different mapping pipelines from imagery to perception:* In the *Nature* study, a decoder was added to the CNN classifier. It is worth experimenting with other translation techniques such as CycleGAN.
- *Imagery is distorted (scaled translated) compared to perception:* Even though the 2 sets of eye gaze data are similar, retrieval is difficult because the recall gaze patterns are downscaled and translated, being positioned in a close range around the image center. Providing a reference frame is considered by Wang et al., 2020 and Johansson et al., 2006 to help reduce the distortion, together with the above mapping method. Wang et al., also suggests that instructing participants to imagine during the recall time an upscaled version of the stimuli could reduce the downscaling. This could be useful in a screen-based experiment.
- *Sequence information disregarded in nature study:* Investigating if the sequence of eye movements matters. Gaze patterns provide information on where a participant is directing his/her attention, for how long and in what order. The order was not considered in the *Nature* study since the fixations are similar in encoding and recall but the order is generally different. The sequence information might provide valuable information for the retrieval process.
- *Longer recall time required:* it is possible that other details of the mental imagery (e.g. colour and texture) are hidden in finer gaze patterns, which might require longer recall time. The *Nature* paper only gave 5 seconds. Wang et al., 2020 suggest that the 5s only give enough time for a participant to recall the very basic layout of the stimuli. Gaze patterns partially highlight the visual features a viewer is driven to. Linsley et al., 2019 and Yang et al., 2016 could guide the integration of image content

in future work. Image memorability (Yang et al., 2016 and Isola et al., 2011) could be relevant since the recalled image might reflect the encoded image still present in the episodic memory.

- *Get measures of spatial imagery ability and vividness*: Wang et al., 2020 observed that some participants had more distorted recall gaze patterns than others and therefore the retrieval had a poorer performance for them. Johansson et al., 2006 concluded that people with a good spatial imagery ability can recall images with minimum eye movements and this usually results in scaled, shifted and translated gaze areas. From this perspective, computational retrieval of image content using raw eye movements better suits people with a poorer spatial imagery ability but not a lot of studies quantify spatial imagery ability and vividness in this context.

- *Gather depth information to explore depth in visual imagery compared to perception:* VR facilitates a 3D environment rather than just 2D images on a screen.

Chapter 2

High Level Requirements Specification

The priority of the requirements is emphasised using must (high), should (medium) and may(low).

2.1 Functional Requirements

- The system *must* implement a VR scene adapted from the study.
- The system *must* implement gaze-tracking to detect involuntary eye movements.
- The system *must* accurately classify visually imagined and visually perceived stimuli, based on eye-movement data
- The system *must* provide a frame of reference for recall eye movements.
- The VR environment *should* restrict external distractions during the experiment.
- The system *must* deal with the distortion of recall eye data.
- The VR environment *should* minimise blinking.

2.2 Non-Functional Requirements

- The system *must* run smoothly.
- The system *must* be reliable and allow running the experiment without crashes.

2.3 Interface Requirements

- The VR headset *must* be able to interface with a personal computer to run the experiment and store the data from eye trackers on the PC for later manipulation.

Chapter 3

Project Plan

3.1 Deliverable Deadlines

The deadlines below are marked as milestones in the Gantt chart in Figure 3.1.

1. Friday 6th November 2020: Project Proposal

2. Wednesday 18 November 2020: Deadline for application submission PREC (alternative dates 20 Jan, 17 Feb)

2. Friday 4th December 2020: Literature and Technology Survey

3. Friday 12 February 2021: Demonstration of Progress

4. Friday 30 April 2020: Dissertation

3.2 Gantt Chart

Figure 3.1 displays a Gantt chart created using the ClickUp tool.

If using a VR headset would not be feasible due to health restrictions, the effort put in 'Implement VR Environment' task will be redirected to 'Developing Code' and data collection will be made remotely using appropriate software such as Prolific or MTurk. If other adjustments will be necessary, they will be made throughout the duration of the project.

3.3 Ethics

One of the main risks involved in this study is VR sickness (Gavgani, 2018), but participants will be advised to stop the experiment at any time if they feel unwell. Considering that a large amount of data will be gathered and that focus is required during the recall time,

Figure 3.1: Gantt chart displaying the project plan for VR based project

participants are asked to take breaks. Participants will be briefed prior to the procedure and made aware of all the potential risks involved in using a VR headset. By signing the consent form, participants agree to take part in the test (being aware of the risks) and agree to their data being processed. All participant data will be stored anonymously. After the procedure is finished, participants will be given the chance to ask any questions they may have and will also be debriefed after the experiment ends. There might be a degree of deception involved because telling people their eye movements are being measured tends to make participants act unnaturally by making more, deliberate or even exaggerated eye movements than when they do not know.

Chapter 4

Resources

Some resources listed below might not be necessary depending on how the health restrictions will intervene in the data gathering process.

4.1 Hardware

1. Vive Pro Eye headset (VR based Project)

Description VR headset with integrated gaze-tracker. Required to track eye gaze during recall and encoding activities.

Limitations Access to the headset is necessary throughout most of the development process in order to conduct necessary testing. The health restrictions might slow down the development process.

2. GPU based Computer

Description The VR environment and data processing need to use a lot of computational power.

Limitations GPU based Computer is required throughout most of the development process, testing and experiment.

4.2 Software

1. Visual Studio IDE

Description Required for developing code in Python.

Availability Available for free with student license.

Limitations None

2. Unity Studio (VR based Project)

Description Required for developing the VR environment used for the experiment.

Availability Available for free with student license.

Limitations None

3. Software for remote data collection i.e Prolific, MTurk (screen-based project)

Description Can offer a safe way of data collection.

Limitations Participants need to have gaze-trackers.

4.3 Knowledge

1. Python for ML

Description Programming language to be used for data processing and image discrimination

Availability Plenty of online resources.

Limitations Failure to effectively use classifiers may delay the implementation of the system.

2. Unity (VR based Project)

Description The game engine that will be used for designing the experiment environment.

Availability Crescent Jicol from the Psychology Department has agreed to contribute to the development of VR environment. Online documentation from the Unity website and short courses will be used to learn how to make the best use of the game engine.

Limitations Failure to effectively use the software may delay the implementation of the system. complex.

3. Psychology expertise

Description It is required to design effective virtual environments and experiment procedures.

Availability Holly Wilson from the Computer Science Department has agreed to contribute to the design of the study. Other online resources are also available such as psychology journals.

Limitations Without the validation of the Psychology Community on the procedure used, the experiment results will be unreliable.

4.4 Other

1. Study participants

Description Participants will be required to perform the recall and encoding protocol. They need to be able to perform visual imagery and do not suffer from VR sickness.

Limitations If an insufficient number of participants is gathered the study results can be unreliable. Future health restrictions might reduce the likelihood of people being willing to participate and conducting the data gathering remotely can be very difficult.

Bibliography

BRANDT, S. and STARK, L. (1997), 'Spontaneous eye movements during visual imagery reflect the content of thevisual scene', *Journal of cognitive neuroscience* **9**(1), 27–38.

Ding, Y., Hu, X., Xia, Z., Liu, Y. and Zhang, D. (5555), 'Inter-brain eeg feature extraction and analysis for continuous implicit emotion tagging during video watching', *IEEE Transactions on Affective Computing* (01), 1–1.

Hebb, D. O. (1968), 'Concerning imagery.', *Psychological review* **75**(6), 466.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and Van de Weijer, J. (2011), *Eye tracking: A comprehensive guide to methods and measures*, OUP Oxford.

Johansson, R. (2013), 'Tracking the mind's eye: Eye movements during mental imagery and memory retrieval'.

Johansson, R., Holsanova, J. and Holmqvist, K. (2005), What do eye movements reveal about mental imagery? evidence from visual and verbal elicitations, *in* 'Proceedings of the 27th Cognitive Science conference', Vol. 1054, Citeseer.

Johansson, R., Holsanova, J. and Holmqvist, K. (2006), 'Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness', *Cognitive Science* **30**(6), 1053–1079.

Johansson, R., Holsanova, J., Johansson, M., Dewhurst, R. and Holmqvist, K. (2012), 'Eye movements play an active role when visuospatial information is recalled from memory', *Journal of Vision* **12**(9), 1256–1256.

Kosmyna, N., Lindgren, J. T. and Lécuyer, A. (2018), 'Attending to visual stimuli versus performing visual imagery as a control strategy for eeg-based brain-computer interfaces', *Scientific reports* **8**(1), 1–14.

Kozhevnikov, M., Kosslyn, S. and Shephard, J. (2005), 'Spatial versus object visualizers: A new characterization of visual cognitive style', *Memory & cognition* **33**(4), 710–726.

Laeng, B. and Teodorescu, D.-S. (2002), 'Eye scanpaths during visual imagery reenact those of perception of the same visual scene', *Cognitive Science* **26**(2), 207–231.

Linsley, D., Shiebler, D., Eberhardt, S. and Serre, T. (2018), 'Learning what and where to attend', *arXiv preprint arXiv:1805.08819* .

Mast, F. W. and Kosslyn, S. M. (2002), 'Visual mental images can be ambiguous: Insights from individual differences in spatial transformation abilities', *Cognition* **86**(1), 57–70.

Mastroberardino, S. and Vredeveldt, A. (2014), 'Eye-closure increases children's memory accuracy for visual material', *Frontiers in psychology* **5**, 241.

Pylyshyn, Z. (2003), 'Return of the mental image: are there really pictures in the brain?', *Trends in cognitive sciences* **7**(3), 113–118.

Pylyshyn, Z. W. (2002), 'Mental imagery: In search of a theory', *Behavioral and brain sciences* **25**(2), 157.

Richardson, D. C. and Spivey, M. J. (2000), 'Representation, space and hollywood squares: Looking at things that aren't there anymore', *Cognition* **76**(3), 269–295.

Scholz, A., Klichowicz, A. and Krems, J. F. (2018), 'Covert shifts of attention can account for the functional role of "eye movements to nothing"', *Memory & Cognition* **46**(2), 230–243.

Scholz, A., Mehlhorn, K. and Krems, J. F. (2016), 'Listen up, eye movements play a role in verbal memory retrieval', *Psychological research* **80**(1), 149–158.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T. and Kamitani, Y. (2019), 'End-to-end deep image reconstruction from human brain activity', *Frontiers in Computational Neuroscience* **13**, 21.
URL: *https://www.frontiersin.org/article/10.3389/fncom.2019.00021*

Thielen, J., Bosch, S., van Leeuwen, T., Gerven, M. and Lier, R. (2019), 'Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience', *Scientific Reports* **9**.

van den Boom, M. A., Vansteensel, M. J., Koppeschaar, M. I., Raemaekers, M. A. H. and Ramsey, N. F. (n.d.), *Biomedical Physics & Engineering Express* .

Vredeveldt, A., Tredoux, C. G., Kempen, K. and Nortje, A. (2015), 'Eye remember what happened: Eye-closure improves recall of events but not face recognition', *Applied Cognitive Psychology* **29**(2), 169–180.

Wang, X., Ley, A., Koch, S., Hays, J., Holmqvist, K. and Alexa, M. (2020), 'Computational discrimination between natural images based on gaze during mental imagery', *Scientific Reports* **10**, 13035.

Wang, X., Ley, A., Koch, S., Lindlbauer, D., Hays, J., Holmqvist, K. and Alexa, M. (2019), The mental image revealed by gaze tracking, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–12.

Yang, Z., He, X., Gao, J., Deng, L. and Smola, A. (2016), Stacked attention networks for image question answering, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 21–29.

Image Credits

Figure 1.3.1: Wang et al, 2020 Computational discrimination between natural images based on gaze during mental imagery. Accessed from: `https://www.nature.com/articles/s41598-020-69807-0#ref-CR12`

Figures 1.3.2: Wang et al, 2020 Supplementary Information. Accessed from: `https://www.nature.com/articles/s41598-020-69807-0#ref-CR12`