Prediction of perceived complex visual stimuli based on eye data during perception and visual imagery in a VR environment

Ana Dobre

BSc (Hons) Computer Science
University of Bath
May 2021

Prediction of perceived complex visual stimuli based on eye data during perception and visual imagery in a VR environment

Submitted by: Ana Dobre

COPYRIGHT

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Bachelor of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author. Signed:

Abstract

This project aims to investigate the potential of using encoding and recall eye movements detected in a VR environment as a novel BCI strategy. In order to achieve this, a software that implements the looking at nothing paradigm in VR was created. This project conducted a similar experiment to Wang et al. (2020) in order to computationally retrieve an image based on encoding and recall eye movements and to investigate if this approach can broaden the range of BCI techniques. However, the experiment was set in a virtual reality environment to improve performance of classifiers and reduce interfering perceptual input. The results regarding the quality of the eye data collected using the above mentioned software are not conclusive, therefore the study can not make a definitive statement about the efficiency of the newly proposed BCI.

Contents

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Eamonn O'Neill, for the guidance, availability, and thoughtful critique he has provided throughout this project. In addition, I would like to thank Holly Willson and Christof Lutteroth for their encouragement, availability and insights regarding this project.

I would also like to thank all the participants taking part in this study for making it possible.

Chapter 1

Introduction

1.1 Background

1.1.1 Visual imagery

Johansson (2013) divides mental imagery in 3 levels guided by Marr's framework for understanding information-processing systems. The first level refers to the functionality of mental imagery, what is achieved by engaging in the process of imagery. The second level analyses how the functionality of the third level is achieved. The research concerned with this level studies the inner representation of imagery (Kosslyn et al. (2006)), how these abstractions are processed, and how mental imagery relates to perception (Finke (1989)). The third level refers to the physical layer, the brain, how and where mental imagery emerges in the brain. Starting with the 1970s a large body of research investigated mental imagery and memory retrieval. This project is based on research done on the third and second layer of understanding imagery.

Similar parts of the human brain are activated during visual imagery and visual perception. Object recognition and spatial indexes are processed in the same distinct parts of the brain (ventral and dorsal) in vision and visual imagery. Mental imagery and memory retrieval are deeply related (Wheeler et al. (2000)) and mental imagery is considered to be 'a critical medium for memory retrieval' in Slotnick et al. (2012).

Visual imagery is a cognitive process during which a mental visual representation of a stimulus forms in the human mind without having the stimulus in front of the eyes (Richardson, 1969). Imagery can also recreate other types of stimuli (i.e verbal, tactile). Visual imagery plays important roles in other

cognitive processes such as learning (Boyer, 2008). In the early 1900, strong correlations were found between eye movements during recall and the image formed in the mind (Moore (1903), Perky (1910)).

1.1.2 Looking at nothing paradigm

Studies in the past 25 years reported spontaneous eye movements triggered by the process of visually recalling a stimulus (such as a scene, object etc.) from memory (Brand and Stark (1997), Richardson et al. (2000), Laeng et al. (2014), Scholz et al. (2016)). Same studies investigate the similarities between the eye movements performed during the recall of a stimulus and during encoding the same visual cue. They concluded that strong similarities exist but they are not identical.

A standard experimental design called "looking at nothing paradigm" is used in studies concerned with the spontaneous eye movements performed during recall. This paradigm allows both encoding and recall eye patterns to emerge. Essentially, during the encoding phase, a person is asked to actively process a stimulus (the encoding can be done through seeing or hearing). During the recall phase, the stimulus is taken away and the person is asked to actively think about the stimulus and form a mental representation of it.

In the case of visual imagery, a person is asked to look at an image and analyze it for a certain amount of time (encoding phase) then he/she is asked to form a mental visual representation of the same image while the image can not be seen anymore (recall phase). The eye movements emerging during the encoding and recall phases are tracked using eye trackers.

Previous studies used a large variety of stimuli: rudimentary such as checker grids (Brand and Stark (1997)), complex such as paintings (Johansson (2013)), small or large datasets (Johansson (2013) and Wang et al. (2020) respectively).

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 fixating a cross. Scholz et al. (2014) conducted a similar study using verbal stimuli associated with a specific area in the environment. During recall, in one of the tasks, the gaze of the participants was directed towards the relevant area while in the other task, the eye gaze was directed to an area unrelated to the encoding moment. In both Johansson et al. (2012) and Scholz et al. (2014),

the amplitude of the recall eye movements was higher when the gaze was kept on the area related to the stimulus. Therefore, maintaining the gaze on the same area during encoding and recall leads to more extensive recall patterns. Moreover, both studies suggest that looking at nothing and memory retrieval have a functional relationship. This is the reason why throughout this study the eye movements performed during visual imagery are referred to as recall eye movements.

The main purpose of the looking at nothing paradigm is to allow eye movement during recall to emerge such that they can be further studied and used. The studies implementing the looking at nothing paradigm have a variety of goals: determining the existence of involuntary eye movement during recall (Brand and Stark (1997)), decrypt the recall eye movement's functionality in cognition (Johansson (2013)), quantify the similarities between encoding and recall eye movements (Wang et al. (2020)), computationally retrieve and image from a dataset solely based on recall eye movements (Wang et al. (2020)).

1.1.3 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 the involuntary eye movements during recall reflect the content and spatial layout of an imagined scene that was previously viewed. 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) is a response to Pylyshyn's hypotheses. The paper proposes an experiment that contains two different types of environments: one lit and the other dark. One group of participants are asked during the encoding phase to actively look at an image and then recall it in the two different environments. Another group of participants is asked to perform the

same task in the two different environments but the stimuli used was a verbal description of the same image. The study concluded that imagery could not be strictly linked to spatial indices because participants performed involuntary eye movements of similar amplitudes in both environments. If the Pylyshyn's hypothesis was true, then the dark environment should not allow participants to perform the eye movements as they are not capable of seeing other spatial indices from the environment. 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. Pylyshyn's hypotheses are disregarded in this study as stronger evidence supports the existence of internal images in the human mind and mental representations are not tightly correlated to other spatial indices from the environment.

Johansson et al. (2006) used complex stimuli, but the variation was lacking as they used only one image during the experiment. 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 in a lit environment. 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)). A perfectly dark environment can replicate the recalling with closed eyes and respond to the fact that humans focus better when no other distractions form the environment are impairing their recall capacities.

1.1.4 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), Wang et al. (2020)). 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 mentally 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. The 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.

1.1.5 Image retrieval methods

As the similarities between the recall gaze patterns and the encoding gaze patterns are already established, new functionality can be attributed to the recall patterns. As mentioned above, visual imagery and recall eye movements can be used to computationally retrieve images. This is a novel idea published in 2020 by Wang et al.

However, visual imagery was previously used in the area of Brain-Computer Interfaces (BCI). Kosmyna et al. (2018) developed an EEG-based Brain-Computer Interface 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 was poor (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) aimed to reconstruct perception by developing a deep neural network (DNN) capable of directly mapping brain activity to the perceived stimuli and therefore reconstructing the perceived image from fMRI data generated 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. Since similar brain areas are activated during perception and imagery, this study suggests that visual imagery might also be used as the input for a similar DNN. However, these approaches involve a large number of motor restrictions on participants due to the use of fMRI machines.

In comparison with EEG and fMRI, using the recall eye movements to observe the visual imagery process is a more natural approach, which can be embedded in other systems, giving more freedom to participants. 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 recall eye movements. They concluded that the retrieval is possible and their study is discussed in the next section.

Figure 1.1. The method used when participants were asked to encode and recall an image (Wang et al., 2020)

1.2 Research Problem

1.2.1 Existing solution

Exploiting the similarities between eye movements during encoding and recall, Wang et al. (2020) attempted to computationally discriminate between images based on gaze patterns performed during the recall of those images.

Image retrieval requires high similarities between eye moves of different participants while looking at the same image. In this study, 100 natural stimuli were used (a database large and diverse 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 from 100 other images based on encoding gaze patterns.
- Scenario 2: computationally discriminate a visual image from 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, the looking at nothing paradigm was used. 100 images were encoded and recalled for 5 seconds each, as shown in Figure 1.1. 28 participants took part in this experiment. 200 gaze patterns (half for recall and half for encoding) were gathered from each participant. The gaze patterns were then transformed into 2D density histograms. The study mentions there are strong similarities between encoding and recall gaze data but the distortions in the recall data still persist.

The image retrieval is treated as a classification task where each image represents a class and the input data for the classifier is represented by encoding or recall gaze patterns. K-nearest neighbour 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%). A Convolutional Neural Network was also used for image retrieval in the first two scenarios, following the structure from Wang et al. (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 distortions in recall data, 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, recall gaze movements were matched with encoding gaze movements from the same viewer. The encoding sequences were already mapped to the right image. After this, a recall histogram was assigned to the class of the matching encoding histogram with an accuracy of 66.4%.

Two of the main limitations that lead to poor classification performance were the distortions in recall gaze patterns as shown in Figure 1.2 (the discrimination performance also varies between individuals) and the use of histograms. Encoding the raw gaze data in histograms inevitably leads to the different images but with similar spatial layouts having two similar histograms. Similar spatial layouts attract similar gaze patterns even if the contents of the image refer to different subjects. Even though they experimented with different kinds of histograms, the results were not significantly different. The current approach faces scalability issues, as a larger image dataset would decrease the accuracy further. The distortions in recall gaze data were likely an effect of a missing reference frame and other distractions from the environment. 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).

1.2.2 Research Problem

Traditionally, the looking at nothing paradigm is conducted on a computer monitor or using a projector and a whiteboard in a dark and quiet room. The main drawbacks of the looking at nothing paradigm are the inevitable distractions

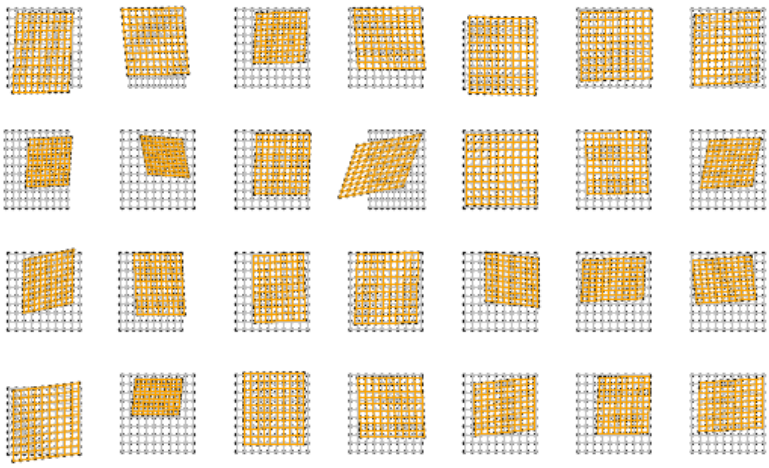


Figure 1.2 Example of distorted gaze patterns during recall (yellow) compared to gaze patterns during encoding (grey). Source: Wang et al. 2020

in the surrounding environment and the lack of reference frames during recall. Even if the room where the experiment is conducted is dark and quiet, participants are still able to pick up visual details from the surrounding environment, as most eye trackers need some sort of lighting source otherwise they would not be able to detect eye movements.

Past studies (Wang et al. (2020), Johansson et al. (2006), (2012)) mentioned the possibility that distortions in recall eye data can be caused by the lack of a reference frame. Implementing a reference frame in the looking at nothing paradigm can lead to improved results in classification as the eye data during recall is expected to hold closer similarities to the encoding eye data.

Only one study aimed in the past to computationally discriminate an image solely based on recall eye movements achieving an accuracy of 72%. In this study, the sequential data (the order in which eye movement occurred together with timestamps) was disregarded. This data can enhance the classification accuracy when certain types of Machine Learning techniques are used (e.g LSTM). As only one study aimed to retrieve images based on recall eye movements, there are numerous other classification pipelines that can be used to potentially increase the accuracy of the retrieval task.

The following Research Questions will be studied in this document:

RQ1 Can an image be computationally retrieved with higher accuracy

than the current state of art based on encoding and recall eye movements?

RQ2 Is the sequence of eye movements during recall and encoding enhancing classification accuracy when trying to computationally retrieve an image?

1.3 Proposed Solution

Virtual Reality technology can provide numerous advantages to the design of looking at nothing paradigm. A Virtual Reality environment is able to replicate a pitch-black room where the only visual stimuli available are the visual cues to be encoded and recalled. VR also gives the researcher great flexibility on the size of the stimuli as well as allowing experimentation with depth and positioning of the stimuli. Creating a software product that implements the looking at nothing paradigm in a VR environment along with other functionalities regarding customizability can later be used as a tool by other researchers. Such a software can be useful as an increasing number of studies regarding the recall eye movements were conducted in the last years. The functionalities that can be attributed to these eye movements have an immense research potential and therefore researchers can benefit from having such a product specialised in gathering recall eye data. Eye-tracking devices can broaden the BCI strategies currently available as eye movements can represent important clues about the human cognition processes. One of the niches that can be of interest for future BCI research is the involuntary eye movements during visual imagery. For such studies, a highly customisable software product that implements the looking at nothing paradigm can represent an important asset.

By using different classification pipelines together with the recall and encoding eye data, the accuracy of retrieving an image might improve. The product described above can be used to conduct an experiment where eye data is gathered and then, in the post-processing stage, different Machine Learning algorithms can be used to perform the classification of encoding and recall eye data. This will test the usability of the software product created while also aiming to improve on the accuracy of retrieving a mental image using recall eye data.

Therefore, this project aims to:

- Create a software that implements the looking at nothing paradigm in VR for maximum minimisation of environmental distractions while also making use of reference frames and providing high customizability levels

for future research in related areas of study.

- Recreate the Wang et al. (2020) experiment and make use of different Machine Learning classifiers in order to improve the image retrieval's accuracy. During the experiment, the eye data gathering will be performed using the above mentioned software. By performing the classification task using this eye data, it can be assessed if the developed software is able to retrieve qualitative gaze data such that the image retrieval process can be performed using it.

1.3.1 Novelty

- Looking at nothing paradigm implemented in VR rather than a computer screen with the environmental surrounding still visible for participants.
- Implementing a frame of reference for the first time in the looking at nothing paradigm.
- The use of time series in the classification process based on recall eye movements.
- The use other ML classifiers than KNN and CNN in retrieving a mental image and evaluate their efficiency

1.3.2 Contribution

The cognitive state of users is something that can be exploited by Human Computer Interaction (HCI) designs using eye movements. Eye-tracking became a technology widely available and researched in the past decade, while eye movements stand at the core of a vast body of research in psychology, but the applications have been lacking. Eye movements can offer clues about the mental processes a person is engaged in (Eger et al. (2007)) and this information can be further used for decoding, guiding and encouraging different types of cognitive processing (Barnes (2008), Liversedge and Findlay (2000), Hayhoe (2017)). Eye movement can be used to determine the intentions of users, the difficulty of tasks etc., and shape the user's interaction with systems (guide attention, affect processing etc). Visual imagery is a part of recalling and it represents a cognitive state that can be tracked and analysed through involuntary eye movements (Wang et al. (2020)). These gaze patterns offer information that can later be used as feedback in a system for a better user experience or as a control strategy.

The involuntary eye movements still raise a lot of questions about their functionality and their possible implementation in a larger software is still to be studied. The only reliable way in which they can be retrieved and analysed so far is using the looking at nothing paradigm. As mentioned before, the most extensive involuntary eye movements during recall are the ones made while looking towards an area related to the encoding phase. Creating a software that implements this experimental approach eases the work of other researchers. The high customizability of such a product allows for different hypotheses to be studied using the same product.

The problem of computationally discriminating a visual stimulus from a dataset based on encoding and recall eye data is worth studying for multiple reasons. Visual imagery can broaden the range of Brain-Computer Interfaces (BCIs) control strategies. Kosmyna et al. (2018) conducted a study aiming to investigate the feasibility of using visual imagery and EEG as a new BCI strategy and explained that 'currently the most common imagery task used in BCI is motor imagery, asking a user to imagine moving a part of the body'. Tan and Nijholt (2010) argue there is too little of a correlation between the user's mental activity and the task carried out by the system, as this lack of logical connection is impacting the performance of the system. It is worth mentioning here that some types of BCI don't work on some individuals. Visual information provides an alternative to individuals who do not have much success with motor or auditory imagery. Using visual imagery would make the process of using BCIs more natural for people. Visual imagery represents a natural process that does not need training and it can represent a natural cognitive strategy for controlling a system. BCIs usually involve fMRI or EEG machines for recording brain signals, which impose numerous motor restrictions on users and the data gathered is often marked by noise. Tracking the recall process of users through eye tracking devices allows for more mobility and wider implementation (important factors in most applications). Besides this, eye tracking is generally a lot cheaper than EEG or fMRI alternatives, far more commercially available, and less training is required to set up and use.

Therefore, visual imagery and eye tracking can represent a simplified and just as efficient BCI strategy compared to the existing ones used nowadays. The simplification is present on both sides, for participants as well as for researchers. A concrete implementation for the retrieval of a mental image using solely eye movements during recall and perception would be a communication software for people with locked-in syndrome that are unable to move their body except for the eyes. The solutions found so far rely mostly on keypads and gaze tracking but implementing the involuntary eye movements during visual imagery would benefit them immensely as the communication with other people would

be performed much faster.

Chapter 2

Literature Technology Review

2.1 Identify different methods for retrieving a mental image

2.1.1 Overview

There is a vast body of literature on fMRI and EEG based classification and reconstruction of perception and to lesser extent imagery. Eye movements were also used in the context of Content Based Image Retrieval (CBIR) as a method of feedback counting towards the relevance of images during search (Zhang et al. (2010), Faro et al. (2010), Hajimirza et al. (2012), Zhou et al. (2018), Coddington et al. (2012)). In the field of Content Based Image Retrieval, eye movements during visualisation are considered to be important cues for image searching but they are usually used together with semantic uses. Most researchers adopt this approach in order to perform the search in massive datasets. At the same time, this kind of approach can not be considered totally transparent and robust. Wang et al. (2020) made the process of image retrieval using eye movement patterns more transparent compared to previous implementations in the field of CBIR. The essential question in Wang et al.'s (2020) study was "how well can images be computationally discriminated from other images based only on gaze data". Even though previous studies demonstrated that encoding and recall eye movements are similar, it was not clear how the eye movements made during visual imagery performed as the input data for computational discrimination of an image. Wang et al. (2020) was the only paper to attempt computational retrieval using gaze patterns from visual imagery.

2.1.2 Brain-computer interfaces

Image retrieval can be tackled as a classification problem where some physiological correlates (such as brain signals or eye movements) represent the input for a computational system that classifies them as a representation of a visual stimulus and therefore an image can be retrieved. This method is used in fields like Brain Computer Interfaces (BCI) where brain signals are used as input, or Human Computer Interaction (HCI) where eye movements can be used as cues for image retrieval.

As Hamamé et al. (2012) explain, 'BCI are widely known for transforming thoughts into actions'. The brain signals used in BCI can be generated using attention (van Gerven and Jensen (2009), van Gerven et al. (2009)), motor intention (e.g. Daly and Wolpaw (2008), Jerbi et al. (2009), Kim et al. (2008), Schalk et al. (2007)) or imagery.

The most used imagery task in BCI is motor imagery. Kosmyna et al. (2018) suggests that visual imagery can broaden the range of BCI strategies. They conducted a study investigating whether using EEGs allows us to distinguish between the mental process of observation and mental imagery of the same visual stimuli for further development of BCI techniques. They concluded that this is possible and highlighted the fact that this approach offers a more natural way of controlling BCIs. Some types of BCI do not work on some individuals - BCI illiteracy. Visual information provides an alternative to individuals who do not have much success with motor or auditory imagery.

There is a vast body of high-quality literature on fMRI-based classification for both perception and visual imagery. Van der Boom et al. (2019), successfully classified visually imagined characters from the early visual cortex. Their accuracy was significantly above chance level using traditional Machine Learning techniques such as Support Vector Machine (SVM). Shen et al. (2019) studied the possibility of perceived image reconstruction from brain activity using a Deep Neural Network (DNN) trained with fMRI data and the images used as stimuli. Their results 'show that the end-to-end model can learn a direct mapping between brain activity and perception'. Reconstruction of imagined visual stimuli has had very little success so far. Shen et al. (2019) tried to reconstruct for visual imagery, but the stimuli used were very simple, and results were rudimentary.

2.1.3 Limitations of BCI

Even though fMRI and EEG based BCIs show impressive results when classifying perception or visual imagery, they impose numerous motor limitations. The lack of mobility imposed on participants by the fMRI machines makes the wide implementation of BCI technologies that use perception or visual imagery extremely difficult. In the case of EEG studies related to BCI, less motor restrictions exist but the setup still poses some issues. However, wireless EEG headsets have been developed and made commercially available in the last years. Another important limitation of BCIs is the lack of logical connections between the user's mental activity and the semantics of the task performed by the system. As Tan and Nijholt (2010) explain, this difficulty in mapping between the two processes can impact the performance of the task severely.

So far, there is strong evidence supporting the fact that during imagery people perform eye movements that resemble the ones made during encoding. This, together with the current gaze tracking technology, could address some of the BCI limitations and offer a more innately efficient solution.

2.1.4 Eye gaze

Eye movements can be used as cues for retrieving a mental image. There are numerous studies investigating optimal techniques for Content Based Image Retrieval (CBIR) and an increasing number of them are making use of eye gaze as a feedback method for the system. Especially in a computational retrieval framework, semantic cues are first used to narrow the search of an image, then the ranked results are further processed using users' feedback for an improvement of the overall performance (Branson et al. (2014), Ribelles et al. (2017)). Wang et al. (2020) suggests a more transparent image retrieval process. Instead of using semantic cues first and then eye gaze only on the already ranked results, they used both encoding and recall eye movements to generalize the computational discrimination of distinct images.

So far, this new approach faces scalability issues. The gaze data is transformed into histograms that encode only the areas within the scene on which the user is fixating and for how long. This means that images with similar layouts but completely different semantic contents can result in similar histograms, making the process of retrieval difficult for large databases (especially when discrete histograms are used. The phenomenon is shown in figure 2.1.

This proved to be an issue even for a database of 100 natural images.

Figure Both images can produce similar eye patterns. The contents of the images are completely different, but their spatial arrangements are very similar.

Despite the inaccuracy of eye movements during mental imagery, histogram similarity still seems to be the main source of confusion for the classification. Possible solutions such as adding more information to histograms will be discussed later. Still, as Wang et al. (2020) suggest, the image retrieval based on eye gaze has improved the number of images that can be distinguished compared to approaches that involved brain signals. Chadwick et al. (2010) and Cowen et al. (2014), both used fMRI measurements and were limited to only three film events and 30 face images respectively as testing data.

Even though the similarities between encoding and recall eye movements suggest promising results in the context of image retrieval, challenges in this field still exist. The two types of eye movements are not identical, as supported by the findings of Johansson and Johansson (2014), Scholz et al. (2016) and Wang et al. (2020) etc. These studies show that gaze patterns observed during recall are distorted compared to those observed during encoding. Moreover, people prefer to recall with closed eyes (Vredeveldt et al. (2015), Mastroberardino and Vredeveldt (2014)), which is unfeasible when video based eye trackers are used in an usual environment (desktop or white board). The next sections will analyse possible solutions for these challenges with the aim of achieving higher accuracy in image retrieval based on gaze patterns.

2.2 Encode and recall eye movements

When focusing on a visual stimulus, humans move their eyes such that the image projected on the retina falls on a specific part of it called the fovea. Here, the cone photoreceptors are concentrated and the highest resolution is achieved therefore, fine details can be observed. Eye movements are usually directed

exactly to the stimuli that catch one's attention (known as overt attention) unless covert attention is at play (Holmqvist et al. (2011)).

Mulder et al. (2004) and Jacobson (1932) examined the eye movements during perception and imagery and concluded that the seven extraocular muscles are the only muscles in the human body that show a similar amount of activation during encoding and recall. Therefore, tracking the eye gaze seems like the only option for analyzing imagery using muscle movement.

The eye movements during encoding are overlapping perfectly with the significant objects within a scene and highlight the spatial arrangement of these objects. The eye movements during recall, even though similar to the ones made during encoding, are not perfectly aligned with the object in the scene and fall somewhere in their close proximity (Brandt and Stark (1997), Johansson (2006), Johansson and Johansson (2014), Laeng et al (2014), Richardson and Spivey (2000)). This phenomenon of shifting, translating and scaling of recall eye movements is described as distortion.

Generalising the retrieval process to new images would be extremely difficult without intrinsic similarity between eye movements during encoding and recall. Wang et al. (2020) concluded that the existing similarities are sufficient for a generalised approach to new images but the distortion in the recall gaze patterns needs to be minimised.

2.2.1 Eye movements to be analysed

Schütz et al, (2011) explains that "eye movements are an integral and essential part of our human foveated vision system". Most of the time, studies that use eye trackers in the context of 2D motionless scenes concentrate on fixations and saccades as the data gathered by eye-tracking tools. As Yarbus (1965) and Rayner (1998) explain, these are the essential eye movement behaviours showing the cognitive process performed by participants. Visual perception takes place mainly during fixations, which refer to maintaining the gaze on an object of interest within the scene. They usually last for 200-300 ms. A saccade represents the rapid eye movements made between two fixations when the focus is changed voluntarily and no visual information other than a blur can be perceived. Their trajectory can not be changed once the movement is initialised. The saccades are sudden, and fast ranging between 50 and 100 ms.

In Wang et al. (2020) the eye movements were measured using the main sequence graphs of saccades, the fixation count and duration, and the overall spatial coverage. The recall eye movements contain fewer and longer fixations

than encoding sequences as shown in Johansson et al. (2006) and Johansson Johansson (2014). Same studies proposed that 'retrieval from memory might account for the longer duration of fixations made during mental imagery'.

The fixations during recall do not coincide with the elements' location in the original image. This does not occur during encoding when fixations perfectly overlap the subjects from a visual stimulus. This results in distorted eye movements during recall. Furthermore, only 95% of the recall fixations were located inside of the stimuli domain, while 99% of encoding fixations were within the stimuli boundaries in Wang et al. (2020). These distortions make it difficult to estimate the intended locations from recall fixations alone. More data from the eye gaze can be used to account for these issues.

Time series are additional information that can be retrieved from gaze. Blascheck et al. (2017) presents an analysis of all the eye data that can be retrieved from eye movements. The duration and sequence of eye movements are additional information that can be retrieved from gaze. The sequence information can be added to histograms as scanpats with every joint containing the duration of fixation. The sequence information is disregarded in Wang et al. (2020) explaining "that encoding positions are revisited during recall but the sequence is not reinstated". However, the sequence data might offer additional information when images with similar layouts are considered. The same area in two different photos can represent very different objects that attract different levels of attention. An object that represents the most important feature of an image is most likely at the beginning of an eye sequence, while an object with very little importance but in the same area of the scene will be scanned later in the sequence. Using time series in the classification pipeline can represent one way to better distinguish between different visual stimuli with similar layout and therefore optimising the classification method.

Besides the sequential data, other types of information are lost when simple 2D histograms are used. Wang et al. (2020) lists other types of histograms used: "(1) binary histogram excluding the time spent in each cell; (2) histograms with a third dimension of time; (3) concatenated histograms of differences between consecutive eye positions". No major differences between them were observed. 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. Wang et al. (2020) only gave 5 seconds. They suggest that the 5s only give enough time for a participant to recall the very basic layout of the stimuli. Exploring different recall and encoding times can improve the classification performance, as pictures with similar layouts will be analysed also based on their unique characteristics such as colour of objects and other details. Eye move-

ments sequence can also reflect the image memorability. Gaze patterns partially highlight the visual features a viewer is driven to. Image memorability can be considered in future implementations (Linsley et al. (2019), Yang et al. (2016)). Image memorability could be relevant since the recalled image might reflect the encoded image still present in the episodic memory.

2.3 Virtual Reality as environment and Reference Frames

2.3.1 Overview

Virtual reality can be a promising tool for improving decoding performance for two core reasons: for reducing external perceptual distraction, and for enabling reference frames.

2.3.2 Virtual Reality

An abundance of involuntary eye movements was observed when experiments were based on the looking-at-nothing paradigm (usually an empty screen or a whiteboard) and visual (Brandt and Stark (1997), Johansson (2006). Johansson and Johansson (2014), Laeng et al (2014), Richardson and Spivey (2000)) or verbal (Johansson et al. (2006), Laeng et al. (2014), Richardson and Spivey (2000)) stimuli were used. The study of Johansson et al. (2014) highlights the fact that successful memorisation is increased when the gaze direction is unchanged during encoding and recall (overlap in gaze locations). It is important to minimise the potential interference of covert attention by creating an environment that contains little to no distractions. Such an environment can be facilitated by Virtual Reality (VR).

Vredeveldt et al. (2015) and Mastroberardino and Vredeveldt (2014) highlight the fact that some people prefer recalling with closed eyes. VR can offer a pitch-black environment that simulates shutting the eyes. This will reduce the distractions the participants are exposed to during the experiment. The video-based tracking system integrated in the VR headset will still be able to track the eye movements because a reference frame will be used and thus enough light will be generated for the pupils to be visible.

In order to accurately track eye movements we intend to use the look-at-nothing paradigm since it proved to be efficient in previous studies and it also aligns itself with the necessary condition that the environment lacks distractions.

Also, the direction of eye gaze during encoding and recall needs to stay the same as the recalling performance is proved to be the best when there is an encoding-recall overlap in gaze locations. Therefore, the reference frame will overlap the VR space area where the visual stimuli are presented during encoding.

2.3.3 Reference Frame

So far, the lack of reference frame proved to be problematic in previous studies including Wang et al. (2020). The most noticeable downside was the distortion in recall gaze patterns due to the missing frame of reference. The scaling distortion is sometimes considered to be due to an increased spatial imagery ability (Johanssonet al. (2012), Johansson et al. (2011)). A reference frame can help participants have an accurate indication of how big the recalled image should be and potentially result in recall gaze patterns with weaker distortions, facilitating better computational retrieval.

The reference frame can also be positioned closer in depth to the participant or made larger as long as it covers the same initial area during the encoding. Johansson et al. (2011) showed that poorer spatial ability actually helps the classification task, because participants with these characteristics make more displaced eye movements during recall that better resemble the ones made during encoding. Participants with high spatial ability are not making extensive eye movements resulting in gaze patterns that are concentrated around the center of the original image. Positioning the reference frame closer to the participant can account for the distortion by forcing the eye movements to span over a larger area. Another way to deal with this type of distortion is making the reference frame larger, which can be easily done in a VR environment, and instruct participants to recall an upscaled version of the original stimulus. This may result in distorted gaze patterns for participants with poor spatial ability but this can only be decided after the data analysis. Coping with upscaled versions of eye gaze patterns might be easier compared with downscaled ones. This approach might facilitate better classification of recall eye movements as the gaze patterns become more similar to the encoding ones. A reference frame can also contribute as implicit feedback during recall time and minimise the accumulation of errors over time. A decreased signal-to-noise ratio would be hard to detect during data gathering. Since this study has at its core an imagery task, detecting loss of focus in participants is difficult to account for. Probably the only way to control this situation would also be avoiding distractions and to give participants frequent breaks. As Wnag et al. (2020) explain, “ without an effective approach to increase the signal-to-noise ratio, both training and testing procedures cannot be free from the impact of noisy data”.

2.3.4 Limitations

The minimisation of blinks might be hard in a pitch black environment, as Johansson et al. (2006) suggests. Minimising blinks is desirable because less gaps in gaze data will be formed. So far the issue of blinking does not seem to have a reasonable solution in the context of VR. One of the only measures that can be taken is requesting participants to avoid blinking, and also scheduling enough breaks for them. Enforcing the limitation of blinks is impossible as it is a natural physiological response, therefore the data is still expected to contain gaps caused by blinks.

A downside of the VR headsets is that implementing this technology is probably not feasible in systems that intend to be widely used, but the principle of minimising distractions and the reference frame can be applied to other technologies. This study can be considered a proof of concept regarding the use of looking at nothing paradigm in VR.

Another important limitation of the current VR headsets is the eye tracker's frequency. Even though VR is a technology that gains more importance year by year, the integrated eye trackers are not as performant as stand alone eye trackers available on the market. Compared to previous studies, the eye tracker used in Vive Pro Eye (the VR headset used in this study) has a frequency of detecting the eye movements 6 to 10 times lower than professional eye trackers used in other studies researching the recall eye movements. The frequency of the integrated eye tracker is still good enough to detect fixations during encoding and recall. These are the most important eye patterns as they give the classifier the most amount of data about each class.

VR fatigue is another downside in the current approach. The experiment is not meant to last for a long time as the effort made by the human eye to accommodate in the VR environment, together with the natural loss of focus that is expected to happen during the experiment, are factors that can minimise the accuracy of eye data and error might accumulate over time.

Nevertheless, this project aims to produce a software that integrates the looking at nothing paradigm and it can be considered a proof of concept regarding the use of looking at nothing paradigm in VR. Future improvements of the VR technology can guide future functionalities of this software.

2.4 Classification pipelines

The retrieval process in Wang et al. (2020) is tackled as a classification task where each image represents a class. All the encoding and recalling gaze patterns were separately represented as 2D histograms of various sizes containing 24x24 cells. Each cell contained the duration of fixation on its respective area. Therefore, histograms record where participants look and for how long. 100 natural images were used as stimuli and 200 histograms were computed for each participant.

The following classification methods were used for image retrieval based on eye movements observed during encoding and recalling separately:

- Weighted k nearest neighbour (kNN) using Euclidean distance (classic machine learning classifier and easy to implement) was implemented as a baseline for comparison with k equal to 27. Leave-one-out cross validation was used for the overall accuracy.

- The encoding and recall histograms were given as input to a Convolutional Neural Network (CNN). The structure used for the CNN was similar to the one implemented in Wang et al. (2019) and the results were compared with the kNN classifier.

When the classification was made using only encoding eye movements, the accuracy achieved by kNN was 94.5% and for CNN was 97.5%. This demonstrates that the encoding eye movements contain enough information for computational discrimination between 100 natural images. When the classification was made using only recall eye movements, the accuracy achieved by kNN was 54.3% and for CNN was 69.8%. This poor performance accounts for the impact of spatial distortion in recall movements when compared to encoding movements for the same image.

The retrieval performance in Wang et al. (2020) varies among individuals, as those who actively moved their eyes during mental imagery performed better. The same study further explains that the neural network's accuracy was higher when participants' eye movements during recall resembled the ones during perception. Lower accuracy was observed when eye movements during recall are largely shifted, scaled and translated or when observers do not move their eyes extensively during recall. A better classification pipeline needs to account for these issues.

In order to computationally retrieve/discriminate an image, classic machine learning techniques (SVM, KNN etc.) or deep learning (CNN, LSTM, GAN) can be used. If it is possible to generate input data that is more efficient

for image retrieval, classic machine learning techniques might potentially be extremely effective. As shown in Wang et al. (2020), the eye movements during encoding offer enough data for a simple classification method such as kNN to perform effectively. If the data fed to the classifier will still be affected by distortion, more complicated machine learning techniques need to be implemented to cope with the distortions and accurately classify the recall gaze patterns.

The performance of classification can be improved by modifying the data used as input (i.e histograms) but also by using different types of classifiers. One other deep learning technique that might prove to be useful is a Long Short Term Memory (LSTM) Network. Tirrupatur et al. (2018) conducted a EEG based study and used LSTM for classifying the state of the human mind and achieved an accuracy of 40%. They used the temporal data generated by the EEG as time series fed to the LSTM. This technique proves to be useful in the context of our project as LSTMs are a type of classification algorithm that enables sequential information to be retained.

2.5 Distortion in recall eye movements and coping with it - mapping encode and recall eye movements

As stated before, encoding and recall eye movements are similar but not identical. Distortion occurs in the recall gaze patterns as scaling, shifting and translation, due to lack of reference frame. This phenomenon is problematic because the classification accuracy decreases as the distortions increase. For an accurate classification, these distortions need to be minimised. Wang et al. (2020) addresses this issue by mapping the recall histograms to the encoding ones.

The remaining similarities between the two types of gaze patterns still allow for a mapping between the two sequences that can improve the classification for recall eye movements. A second task was added to the CNN, forcing it to explicitly learn the mapping between encoding and recall data. The classification network also contains a decoder that generates the associated encoding histogram based on a recall histogram. Both of these data abstractions refer to the same stimulus. The leave-one-out test achieved an improved 72.1% accuracy performance. Wang et al. (2020) explains that they “did not find any consistent distortion patterns among the recall eye movements even for one dataset of one observer”. This can be the reason for such a low improvement when using the explicit learning of the mapping between recall and encoding. Better performance can be achieved by implementing other mapping pipelines.

Generative Adversarial Networks (GANs) can implement image-to-image transformation when two sets of images are available and mapping between them is advantageous. Generally, there are two types of image-to-image transformations: paired and unpaired. Paired training samples are difficult to be obtained, which might be the case in this study considering the inconsistent distortions in recall eye movements even for one dataset of one observer, while unpaired training samples can be advantageous but the risk is that less satisfactory results are produced (Tripathy et al. (2018)). In our case, the image-to-image transformation refers to recall-encoding transformation. One method that uses unpaired training samples is CycleGAN. This GAN architecture uses unsupervised image translation models that can improve the mapping approach used by Wang et al. (2020). Even though more options are available for paired recall-encoding transformation, they require large datasets of paired images that can be difficult to compute in our context due to ample distortions between encoding and recall eye movements. Pix2Pix GAN is an approach that implements paired transformations and it might produce good results if the histograms of recall eye movements are manipulated to be more similar to encoding histograms.

Other machine learning techniques can be implemented for the mapping problem as long as the recall histograms incode less distortion. Some of the techniques that can be implemented are Scale Invariant Feature Transform (SIFT) (Karami et al. 2017), Speeded Up Robust Features (SURF) (Bay et al. 2006), Features from Accelerated Segment Test (FAST) (Rosten et al. 2006), Hough transforms (Goldenshluger et al. 2004) or Geometric hashing (Tsai 1994).

Essentially, the purpose of implementing these mapping techniques is to minimise the impact of the distortions in recall eye movements for the classification task. In the alternative scenario in which distortions are dealt with through the Virtual Reality environment and new histograms are designed to retain more data about the eye movements, mapping might not be necessary.

2.6 Conclusion

A few possible solutions to the problems faced by Wang et al. (2020) were listed above. Until implementation and testing of the methodological approach, the issues stated above still generate the following research questions:

RQ1 Can an image be computationally retrieved with high accuracy based on encoding and recall eye movements?

RQ2 Is the sequence of eye movements during recall and encoding

enhancing classification accuracy when trying to computationally retrieve an image?

Chapter 3

Design

3.1 Overview

This chapter aims to describe the design of the software implementing the looking at nothing paradigm in VR. The software aims to be suitable for other researchers that are studying the spontaneous eye movements produced during the recall of a visual stimulus. The software implements both the encoding and the recall task in a Virtual Reality environment that minimises distractions and makes use of reference frames.

The visual stimuli used are images depicting outdoor and indoor real scenes. The eye data gathered from users is saved in a .csv file together with timestamps, and other variables that help distinguish between sessions and tasks.

3.2 Operating environment

Initially, the software was meant to be used in a laboratory, where setup conditions (such as position of the base stations, eye calibration, participant position in VR etc.) could be controlled by the researcher. Besides this, the researcher was meant to control the frequency of performing tasks and breaks as well as dealing with the various settings that need to be handled at the start of each session with a new participant.

Due to the current health restrictions imposed by Covid-19, the software had to be deployed to the participant's homes. This implies that the software needs to integrate a comprehensive interface for all types of users and detailed instructions about the VR setup, installer and usage of the software had

to be provided to the participants such that they are guided through all the steps of the experiment. New requirements need to be created to accommodate these changes such as remote data collection and flexibility regarding the location of the experiment.

3.3 Requirements

3.3.1 Overview

The following section will describe the Requirements needed to build the product software that is implementing the looking at nothing paradigm in Virtual Reality. The design of the system is based on the functionalities outlined below.

3.3.2 Requirements Elicitation

Multiple sources were used to generate the requirements. The Literature Review was the main source of information for requirements elicitation. The overall layout of the looking at nothing paradigm is described in section 1.1.2, while the overall characteristics of the VR environment are analysed in section 2.3. The rationality behind each decision regarding the visual stimuli and reference frames are detailed in the sections 3.4 and 3.5.

Some requirements are introduced such that the suggested aims and novelties, discussed in chapter 1, could be achieved. Examples of such requirements include the usage of virtual reality and longer recall time.

3.3.3 Prioritisation

The priority of each requirement is considered using a three level scale (High, Medium, Low). The decision regarding the priority is made based on how critical is the implementation of each requirement for the aims of this project.

3.3.4 Functional Requirements

1. System logic 1.1 The system must implement the looking at nothing paradigm in VR. - High

1.2 The system must be able to retrieve all eye data generated during Recall. - High

1.3 The system must be able to retrieve all eye data generated during Encoding. - High

1.4 The system must check if the user is looking at the image during the Encoding task. - High

1.5 The system must check if the user is looking inside the reference frame during the Recall task - High

1.6 The user must be able to pause the system at any time. - Medium

1.7 Frequent breaks must be scheduled between tasks. - High

1.7 The system must detect if the user is blinking. - High

1.8 The system must display a visual noise mask between each task. - High

1.9 The system must display a fixation cross before each stimulus to redirect the users gaze towards the center of each image. - High

1.10 The system must display a countdown before each tier to allow the user to prepare for starting the encoding task. - Medium

1.11 The system must allow the user to complete 3 sessions each of 100 images so that enough data is retrieved for the classification task. - Medium

1.12 Each session must be split in 10 tiers so that participants do not lose focus and avoid VR sickness. - High

1.13 The system must automatically redirect the user to the menu after a set of tasks is completed. - High

1.14 The system must provide the user with a way to access the forms. - Medium

1.15 The system must provide the user with a practice session before starting the official study so that participants get used to the tasks. - High

1.16 The frame of reference should be visible during recall and encoding. - High

1.17 The user must be provided with a survey that assesses their visual imagery vividness capabilities. - High

1.18 The user must be provided with a survey that assesses their spatial memory capabilities. - High

1.19 The user needs to finish one session before starting a new Session such that all the images in the set are viewed once per session. - High

1.20 Everything must be in the Field Of View (FOV) of the user. - High

1.21 The seated user must see the stimuli and frame right in the middle of their FOV. - High

1.22 Users must be provided with a sign on the ground that represents where they should stay during the experiment. - High

1.23 The encoding task must last 5s exactly. - High

1.24 The recall task must last 8s exactly. - High

2. Sequence of events

2.1 The sequence of events throughout the study should follow the following structure: - High

1.The user enters the participant ID

2.The user completes a Vividness of Visual Imagery Questionnaire (VVIG)

3.The user completes a practice trial to familiarise with the task.

4.The user starts a tier of 10 images for the first session by pressing the “Start tier for Session 1”

5.Display a 3 seconds Countdown before a tier starts.

6.Constantly display the frame of reference without changing its size or position throughout the encoding and recall task.

a. Display the fixation cross for 1 second.

b. Display for 5 seconds a randomly chosen visual stimulus for the encoding task.

c. Display a dynamic visual noise mask for 0.5 seconds.

d. Remove anything else but the reference frame for 8 seconds during the recall task.

7. Repeat steps a to d ten times until one tier is done.

8. The user is redirected to the menu and they can take a break.

9. A session is completed after all 100 images are viewed (10 tiers).

10. The user can start a new session only after the previous session is completed so all 100 images are displayed once in a random order.

11. The user repeats steps 4 to 10 for each session.

12. The user completes a spatial memory survey at the end to avoid influencing their behaviour during the study and to assess their spatial memory capabilities.

3. Visual Stimuli

3.1 The data set needs to contain enough stimuli for the classification task to be achieved. - High

3.2 The stimuli need to be complex enough so that the gaze data retrieved contains enough information for the classification task. - High

3.3 The stimuli dataset needs to be diverse. - High

3.4 Distance between user and image needs to be comfortable for the human eye in a VR environment. -High

3.5 Size of the stimuli needs to occupy 20 degrees in the field of view. - High

3.6 Each stimulus from the dataset needs to have the same resolution. High

3.7 The light conditions of the images need to allow the user to perceive details. - High

3.8 The display time of the visual stimuli needs to be 5s. - High

3.9 The images are displayed in random order to minimise biases. - High

3.10 The image needs to be located in the middle of the user's field of view. - High

4. Reference Frame

4.1 The reference frame needs to minimise the distortions in the recall gaze patterns. - High

4.2 The frame needs to dynamically reshape depending on the size of the displayed stimulus in order to perfectly frame the stimulus displayed. - High

4.3 The frame needs to be positioned in the same location as the displayed stimulus such that it creates a border around it. - High

4.4 The reference frame needs to not have a salient effect on the user. - High

4.5 The reference frame needs to emit enough light so that the eye

trackers can detect eye movements even within an extremely dark environment. - High

4.6 The Frame needs to be visible during the whole experiment - High

4.7 The frame needs to be the only visible object during the recall task for 8 seconds. - High

5. Environment

5.1 The environment during the study needs to be pitch black and to lack distractions. - High 5.2 Provide a sign for where the user should stand. - High

5.3 The background colors need to delimitate the encoding and recall tasks from the navigation of the system through the menus. - Medium

5.4 The environment during the study still needs to contain enough light for the eye tracker to work and for the sign on the ground representing the ideal position of the user to be visible. - High

5.5 Performance of the Unity environment needs to be high in order to minimise low frame rate that might affect the eye tracker frequency. - High

6. Gaze tracking

6.1 The system must make use of the eye tracking for the encoding and recall task. - High

6.2 The system must collect eye data throughout all tiers and sessions. - High

6.3 The system must log all the raw gaze data. - High

6.4 The system must check for blinks. - High

6.5 The system must detect if the user is looking at the stimuli or frame or somewhere else in the environment. - High

6.6 Retrieve eye data with highest accuracy possible. - High

7. Customizability

7.1 The system must be customizable for other researchers to use it according to their needs. - High

Change stimuli set.

Change the number of sessions.

Change the number of images in a tier.

The duration that the stimulus is visible for. The duration that the reference frame is visible for.

Distance between users and stimuli.

Change stimuli size.

The reference frame must dynamically change its shape depending on the size of each stimulus even if the data set chosen by a future researcher contains stimulus of various sizes.

The data collection needs to be done remotely.

8.Hardware Requirements

8.1 It must be possible to use the Vive Pro Eye virtual reality headset with the system. - High

8.2 SRanipal needs to be installed on the computer and the Vive Pro Eyes built-in gaze tracker must be usable throughout the whole experiment. - High

8.3 The user must be able to use the Vive Controller to press buttons and pause sessions. - High

3.3.5 Non-Functional Requirements

9.Non-Functional Requirements

9.1 The system must function reliably, with no performance drops given that the hardware requirements of Virtual Reality (sufficiently powerful CPU/GPU) are met. - High

9.2 The system must be highly customizable, in order to be reused for other similar studies and to comply with the needs of other researchers. - High

9.3 The gaze data must be accurate and be collected only when the user is looking at a visual stimulus or recalling the previously seen image. - High

9.4 The system must be safe to use with adults. - High

9.5 The system must not induce VR sickness on users. - High

9.6 The system architecture must be maintainable and extendable. - High

9.7 For each participant 400MB of free memory is needed. - High

9.8 The timing of retrieving gaze data must be extremely accurate so that the eye data retrieved for the encoding task does not include eye movements performed during visualising the fixation cross or noise mask. Same applies for the recall task. - High

3.4 Visual Stimuli

3.4.1 Overview

This section will explain the process of choosing the stimuli database and all aspects regarding them such as size, position in the Virtual Reality environment etc.

3.4.2 Content and Resolution

Similar studies conducted in the past made various decisions regarding the stimuli used. Brand and Stark (1997), one of the first research papers investigating the involuntary eye movements during visual imagery, used simple visual stimuli such as a 6x6 grid with random cells coloured in black or whilte. Johannson et al. (2006) used one of Nordqvist's paintings from 'Ruckus in the Garden' as the only stimulus, but this approach lacks variation even though a complex stimulus is used. Wang et al. (2020) is the only research so far to use multiple complex stimuli. Their dataset contained 100 images of various sizes and resolutions representing indoor and outdoor natural scenes.

Using complex stimuli is a desirable approach in the context of this study. First of all, close similarities must exist between the stimuli used in order to accurately compare the classification accuracy of this study to Wang's results. Therefore the image database needs to contain complex indoor and outdoor natural scenes so that similar amounts of eye data can be retrieved during the recall and encoding tasks. Second of all, using such a dataset gives a better perspective on the potential of using eye movements as a new BCI technique. Moreover, encoded and recalled complex scenes produce enough data for the classification algorithms, more rudimentary stimuli will not produce enough gaze points.

Multiple datasets of various sizes from CVonline, Kaggle and ImageNet were considered during the design process. Initially, due to the uncertainties created by the Covid-19 pandemic, a solution for a potential small number of participants had to be found. In order to perform the classification task, around

6000 eye patterns are needed, half for recall and half for encoding. Five participants was the smallest possible number of participants considered as baseline. In this case 600 images are needed to reach the number of eye patterns mentioned above. To create this dataset, the 100 pictures used in Wang et al. 2020 and another 500 from Judd et al. (2009) were used. The later dataset contains over 1000 images used by Judd et al. to predict where humans look when provided with complex visual stimuli. The 100 images used in Wang et al.'s study are also part of Judd et al.'s dataset. The latter contains images of a wide variety of subjects, lightning, resolution, and camera angles. The criteria used to choose the 500 images were the highest resolution possible, clear and unblurred images, similar lighting and perspective, unedited and normal saturation, images depicting real scenes, no dim or poorly lit scenes so the users clearly understand what they see, similar camera angle. The recommended image format for Unity is .jpeg so datasets of this format were considered.

Another approach considered was to use only 100 images viewed multiple times to create the set of gaze patterns needed for classification. In this case, the 100 images from a session and the number of sessions can be easily scaled depending on the number of participants available. This approach raises questions about the differences in perception and visual imagery for an image viewed for the first time and the same image being viewed for the second or third time. Even though the cognitive process is different between sessions, because participants are getting more familiar with the images each time they are displayed again, participants get a chance to look at the image in closer detail, integrating different details each time (as participants also mentioned). Seeing more/other details each time is also due to the familiarity and short encoding time. This creates a wider set of valid eye movements for a single image, potentially leading to better classification capabilities. For example when an image depicts a complex scene some participants may be looking at the same main subjects within the photography but the secondary details they decide to focus on can be different. Viewing the same image multiple times can even this bias as there is a finite number of details and salient objects a person is likely to fixate during encoding and therefore visually imagine in the recall phase. This is the approach used in the experiment, the software will include 3 sessions each containing the same 100 images.

The 100 images used were chosen based on the same criteria mentioned above, but all images had to have the same resolution. Wang et al. used both landscape and portrait images with various resolutions, thus facilitating the classification process. In this study it was decided to use only pictures with a resolution of 1024x768, as this is the highest and predominant resolution in Judd et al. and Wang et al. 26 pictures from Wang's dataset were passing the criteria

of resolution and clarity. Another 48 images were selected from Judd et al. To complete the set of images, 32 images from Hosu et al. 2020 were selected. The last mentioned dataset represents a database with over 10 thousand 1024x768 pictures, however very few of them were passing the criteria mentioned above for selecting appropriate images for this study. Higher resolution for the stimuli is desirable but no image dataset large enough and that was passing all the criteria (complex real scenes, landscape images of the same size etc.) was found.

3.4.3 Size and Position

The size, position and resolution of the visual stimuli are particularly important in the context of this study. The appearance and location of the images can change the perception and understanding of the participants about what they see and therefore what they are able to recall. The size and resolution of the stimuli influence the level of detail perceived by humans through the visual system. The position of the stimuli changes the convergence distance of the eyes and the position of the focal point, generating differences in the amount of details encoded by humans. The decisions made regarding the size, position and resolution need to facilitate the visual perception process and lower the effort of participants to property integrate what they see. In an ideal scenario, participants should not struggle to understand the contents of any stimulus and this needs to be addressed with special care considering the usage of VR headsets.

The Headset specifications published by HTC are as follows:

Screen: Dual OLED 3.5 diagonal

Resolution: 1440 x 1600 pixels per eye (2880 x 1600 pixels combined)

Refresh rate: 90 Hz

Field of view: 110 degrees

Together, these specifications offer clues about the capacity of Vive Pro Eye Headset to render details in a VR environment. Even though the resolution of the screen is high, due to the magnifying effect of the lenses necessary in any VR headset, the individual pixels forming the screen can still be picked up by the human eye. Therefore, the image formed through the VR headset does not resemble the normal 20/20 vision of the human eye. These effects need to be minimised through the size, location and resolution of the visual stimuli. The integrated eye tracker specifications also need to be considered in the design

process:

Gaze data output frequency (binocular): 120Hz

Accuracy: 0.5°-1.1° (Within FOV 20°)

Calibration: 5-point

Trackable field of view: 110° (Eye surgery, eye disease, heavy makeup, and high myopia may affect in eye tracking performance)

Data output (for each eye): Timestamp (device and system), Gaze origin, Gaze direction, Pupil position, Absolute pupil size, Eye openness

Interface: HTC SRanipal SDK

3D engine compatibility: Unity, Unreal WorldViz Vizard

These specifications indicate the accuracy of the retrieved eye data with the Vive integrated eye tracker. Aiming for the highest accuracy of gaze data and highest frequency available from the eye tracker is particularly important in this experiment as mentioned in the Requirements. The accuracy and frequency of the eye date both have an impact on the classification task since they determine the preciseness of the eye dataset.

The eye tracker has the best accuracy within a field of view of 20 degrees, as the HTC specifications above indicate. The visual stimuli are the only objects of interest during the experiment and therefore most of the gaze positions produced by users will fall within the area of a stimulus. This means that the visual stimuli need to occupy roughly an area of 20 degrees in the participant's field of view such that the eye tracker can accurately detect participants' eye movements.

One degree in the field of view can be determined using the following formula: a = tan(1) x d

Where:

a = area in cm for one degree in FOV

d = distance between participant and stimulus

NB: In order to solve the calculations, the distance from the participant to the stimuli needs to be decided first.

Position on the Z axis: In the context of previous studies, the position of the stimuli was usually recommended by the eye tracker manufacturer. Remote eye trackers were used in similar previous studies and their specifications mention the ideal distance from participant to eye tracker and therefore ideal distance

from participant to screen (remote eye trackers are placed under the computer's screen). The usual distance used in previous studies is between 0.6 m and 0.7 m. However, the eye tracker used for this study is integrated in the VR headset and its specification does not include any restrictions about the ideal distance between user and stimuli. The manufacturer of the VR headset (HTC), as well as the documentation for the engine used to create the experiment's environment (Unity), recommended that the comfortable distance between the user and objects in the 3D virtual space is between 0.75 m and 3.5 m. The chosen distance to position the stimuli in the virtual environment was 0.75 m in front of the participants, as it is the closest acceptable distance, in a VR environment, to the distance used by Wang et al. (2020) (where 0.7 m was used). It is desirable that the distances are similar so the distance can not be counted as a confounding variable.

Therefore, in the case of this study, where the distance from eye to stimulus is 75 cm, one degree in the field of view represents 1.4 cm (using the above formula). Since the maximum size of the displayed stimuli is 20 degrees, and the images are displayed right in the middle of participant's FOV, the maximum width of an image should be 28cm. This will ensure the highest accuracy possible for the eye data gathered during the experiment.

Position on the X and Y axes: Since the experiment has a long duration, the environment was designed for participants to sit down. The average eye level of a person while sitting down in this position is 1m. The centre of the canvas displaying the stimuli is positioned at 1m on the y axis and at 0 on the x-axes so that all stimuli are in the middle of the field of view of the participants as long as the participants sits at position (0,0,0) which is indicated to them using a sign in VR.

3.4.4 Duration of the encoding and recall tasks

In normal lightning conditions (i.e. when the human eye does not struggle to interpret the data from the surrounding environment) a visual stimulus is registered after 80ms. A normal human is able to perceive and interpret a complex stimulus such as an image in 150ms, as explained in Rayner et al. (2009).

The encoding task will last 5s during the experiment. As shown in Wang et al. (2020) the eye movements generated in 5s during the encoding task contain enough information even for simple classifiers such as KNN to perform well.

Figure 2.2 Amplitude of encoding and recalling eye movements. SOurce: Wang et al. 2020

Longer recall time, of 8s, is implemented for the recall task. As explained before, the eye tracker has a lower frequency of detecting eye data. As the recall gaze patterns are not perfect replications of the encoding gaze patterns, the classification task might benefit from longer sequences of data in which users have enough time to remember all the visual information processed during the encoding task. Moreover, as mentioned before, information such as colour and texture might appear in the recall gaze patterns later on in the sequence, as people tend to firstly remember the spatial layout of an image.

3.5 Reference frame

3.5.1 Overview

Previous studies investigating the eye movements that emerged during visual imagery of a stimulus concluded that strong similarities exist between encoding and recall eye movements (Brandt Stark 1997, Johansson et al. 2006, Johansson et al. 2011, Johansson et al. 2012, Gbadamosi Zangemeister 2001, Wang et al. 2020). The same studies consistently reported distortions in the recall eye movement compared to the encoding eye movements and one possibility why these transformations appear, as suggested in these papers, is the lack of a reference presented during recall time. The distortions include downscaling, which will be discussed in detail in the 'Size' section, translation, skewing, stretching and even uniform scaling on the same dimension (such as a fish-eye lens effect). The spatial distribution of fixations during imagery becomes smaller and crowded towards the centre of the recall area as seen in Fig 2.2. Therefore the distortions can vary, making the process of classification difficult.

3.5.2 Importance and positive effect

Reference frames are used to encode spatial information from the surrounding environment. As Klatzky (1998) and Wang (2007) explain, any reference frame can be defined using a reference point and the direction of the axes. When a frame is defined based on the reference point, a person can encode the surroundings based on an egocentric relationship (object-to-self) or an allocentric relationship (object-to-object) as Wang (2012) explains.

Egocentric and allocentric are two spatial representations that help people remember visually and spatially information from the outer world. They are concerned with the spatial arrangement of the viewer in relation to objects in the world and positioning relations between objects. In the case of this study, the reference frame can be considered allocentric during encoding and egocentric during recall, enhancing the cognitive processes of perception and visual imagery.

3.5.3 Position

Johansson (2013) explains that recalling in the same area is beneficial for involuntary eye movements to occur. Allowing participants to recall while looking at the same area where encoding eye movements are spanned enhances the amplitude of eye movements. The position of the reference frame is exactly where the previously seen image was before.

The frame of reference will constantly be displayed both during the encoding and recall to allow for continuity in participants' perception of the environment and show the direct relationship between frame and stimuli. At the same time, the frame must not be considered a novelty during recall so it does not attract the gaze. The participant gets familiar with the size, shape and aspect of the frame during the viewing of the fixation cross, stimulus and noise mask.

3.5.4 Appearance

The aspect of the reference frame needs to be considered carefully since it is the only visible object during the recall time. The attention of the user should be focused on the interior of the frame, not the frame itself, therefore the salient effect needs to be minimised.

Considering that the frame is the only visible object during the imagery task it is easy for participants to get distracted if the frame stands out in the

Figure 2.3 Resizing of frame

completely dark environment. At the same time, the design of the experiment still needs to follow the looking at nothing paradigm as this model is considered to encourage the recall eye movement in participants the most (Johansson 2013).

So the main aspects to be considered are the thickness, colour and luminescence of the frame. To minimise the salient effect, the frame was made thin enough to still be perceptible during the encoding phase, so the participants are aware of its existence even when the bright colours of the stimuli are receiving most of the focus and the continuity effect between encoding and recall is maintained. This means that the allocentric contribution of the frame is not diminished. The decided colour of the frame is white. This colour is the least contrasting one no matter the colours contained in the displayed image. The luminescence of the frame was changed compared to the rest of the objects in the environment resulting in a light grey. Dimming the radiance of a frame further reduces the allocentric perspective the frame can give to the environment as reported by Ruotolo et al (2011).

The result is that the frame of reference is still distinguishable and achieves the original purpose of guiding the recall process by constantly making the participants aware of the initial size and position of the stimuli. It can be argued that looking at nothing paradigm is still respected using such a design. Additionally, other studies were not using completely empty environments, the "nothing" was still containing the whiteboard edges or the computer screen edges (but they can not be considered a reference frame as they were not giving exact information of where the stimulus was before).

3.5.5 Size

During the design process, enhancing the size of the reference frame was considered in light of previous studies reporting downscaled gaze patterns during

visual imagery. By enhancing the reference frame during the recall time and asking participants to imagine an upscaled version of the image, the aspect of the distortion that refers to downscaling can be diminished. At the same time, asking people to imagine an upscaled version of a stimulus can enhance the amount of detail contained by the mental image (Hubbard and Baird (1988), Kosslyn (1978)). It is hard to argue that this would be the case here. The previous study that mentioned this was asking participants to create a mental image after seeing a word. A lot of freedom was given to participants, during that experiment, since they were allowed to 'create' a new image and not to think of a fixed image as in the case of this study. All the details that should be imagined are contained within the stimuli, allowing people to manufacture a similar but slightly different image of a fixed stimulus is not ideal in this case. The eye movement describing the recalled image needs to refer to the contents of the image viewed and processed during encoding. A more detailed representation of the picture is meant to be achieved by a longer recall time (which will be discussed later) not by asking for an upscaled version. Another uncertainty raised by imagining an upscaled version of an image is the accuracy of scanning during imagination. This approach has a chance of still producing distortions such as translation, skewing, stretching and so on.

The data reported by previous studies were not giving explicit indications about distortions (such as percentages of downscaling and other types of transformations) nor how many people are making distorted enough eye movements during recall to negatively affect the classification process. The Wang et al. 2020 study explains that 'We did not find any consistent distortion patterns among the recall eye movements even for one dataset of one observer'. Another study conducted by Brand and Stark (1997) reported a 20% downscale in the recall gaze patterns but the stimuli used were lacking complexity as they were represented by an 8x8 grid with random cells coloured in black or white. Consistent findings suggest that the downscaling is accentuated in people with high spatial memory capabilities, so not all people are subjected to downscaled versions of recall eye movements. Testing the participants for this characteristic and then proceeding with a design that integrates this specific feature for each participant is beyond the purpose of this study. Therefore, the data referring to distortions available at this point is not conclusive about occurrence, severity and patterns, so the problem of classification becomes nonlinear and more complex. Just by enhancing the size of the reference frame, the recall eye movements might not get significantly more similar to the encoding gaze patterns. An advantageous effect of the upscaling might be that the current problem might become the opposite one, where people with a good spatial memory perform gaze patterns similar to the encoding ones, while people with poor spatial abil-

ities just create upscaled versions. The benefit in this situation might be that it would be easier to classify more dispersed fixations than crowded ones, as the crowded ones have a higher probability to be mapped to the wrong encoding equivalent fixation. Some preliminary studies need to be performed for this hypothesis. Considering that the distortion patterns were not consistent even within the dataset of the same participant, the distortions can be influenced also by the content of the stimulus and the specific response of a participant to that image.

An upscaled reference frame makes sense in the context of an upscaled imagined representation of a stimulus so it is necessary to instruct participants to create an upscaled version of the previously encoded image. The task of recalling a larger image than the one encoded implies different cognitive processes than just imagining the exact perceived stimulus as Kosslyn (1978) explains. The efficacy of the reference frame can not be completely established in such conditions. Considering the number of participants available during current health restrictions and the large number of trials needed for classification, creating a model that tests for each condition separately necessitates more time than it is available.

An interesting approach would be creating a visual illusion where participants perceive the image as being larger than it is in reality. The famous Delboeuf illusion (Gentaz Hatwell (2004)) was considered for this approach. Unfortunately, for most illusions that are concerned with size, a form of comparison is needed otherwise no illusion occurs. Displaying 2 images, one inside the frame and one outside, in order to achieve the illusion, would be confusing and distracting, effects that are highly undesired given the nature of the task.

Given the uncertainty provided by how people would react to an enhanced reference frame and the task of imagining a larger version of the initial stimulus, it was decided to keep the size of the reference frame the same as the stimulus both during recall and encoding and ask participants to imagine the stimulus exactly as it was perceived. No other study researching the eye movements during recall used reference frames.

3.6 VR Environment

3.6.1 Diagram of objects in the scene

As described in sections 3.4 and 3.5, the stimuli and frame overlap and they are both positioned 0.75m in front of the user. The following diagram shows the

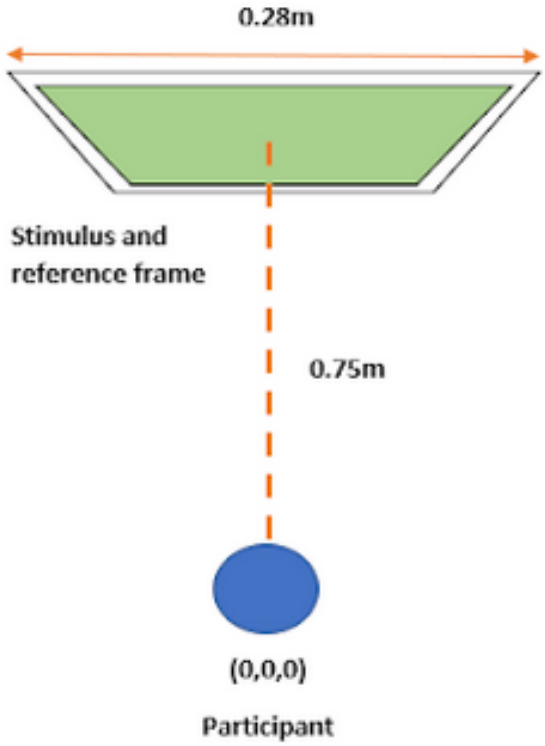


Figure 2.4 View from above describing the positioning of the frame of reference, stimulus and participant inside the VR environment during the encoding phase

spatial arrangement of stimuli, frame and user.

3.6.2 Position of user

The position of the user is particularly important during the experiment. The system permanently displays a green sign on the floor at the position (0,0,0) where the user is instructed to stay during the whole experiment.

3.6.3 Lightning

During the navigation of the system through menus, the environment is white with the line of horison dimly defined. Once the tasks of encoding and recall start, a countdown of 3 seconds is suggesting the users that they should prepare to start the experiment and the environment turns pitch-black in order to min-

imise distractions. The only objects of interest within the field of view of the user are the stimuli and reference frame. This eliminates all visual distractions that may exist during the experiment.

3.7 Forms

The user needs to complete 2 forms that assess his/her vividness of visual imagery and spatial memory capabilities. The two need to be investigated separately as people seem to encode imagery information using object imagery and spatial imagery (Kozhevnikov et al. (2010)). People with more developed object imagery tend to encode images as a whole, while people with more developed spatial imagery rely on spatial relationships when encoding images.

Kozhevnikov et al. (2010) identified a tradeoff between the two, that might be related to the fact that both processes used the same limited resources such as attention. As mentioned before, people with better spatial memory make less extensive eye movements during the recall time, leading to more distorted gaze patterns. It is important to access both spatial and object memory capabilities such a more comprehensive results analysis can be performed.

The Vividness of Visual Imagery Questionnaire (VVIQ) created by Marks (1973) was used to assess the object imagery capabilities of a user. The full VVIQ can be found in Appendices. The Spatial Memory test performed by the users represents an adaptation of the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ), created by the Blazhenkova Kozhevnikov in 2009, where only the spatial memory questions were used. The full Spatial memory form can be found in Appendices.

3.8 User Interface (UI)

3.8.1 Overview

Initially the study was meant to take place at University of Bath in controlled conditions. The UI was designed to be handled by the study facilitator. The UI was simple and not self explanatory as the facilitator was considered to be a knowledgeable person about the insights of the study. This version of the software is described in section 3.7.2.

As the health restrictions were not eased, a different, more comprehensive UI version for any type of user and facilitator had to be designed. The

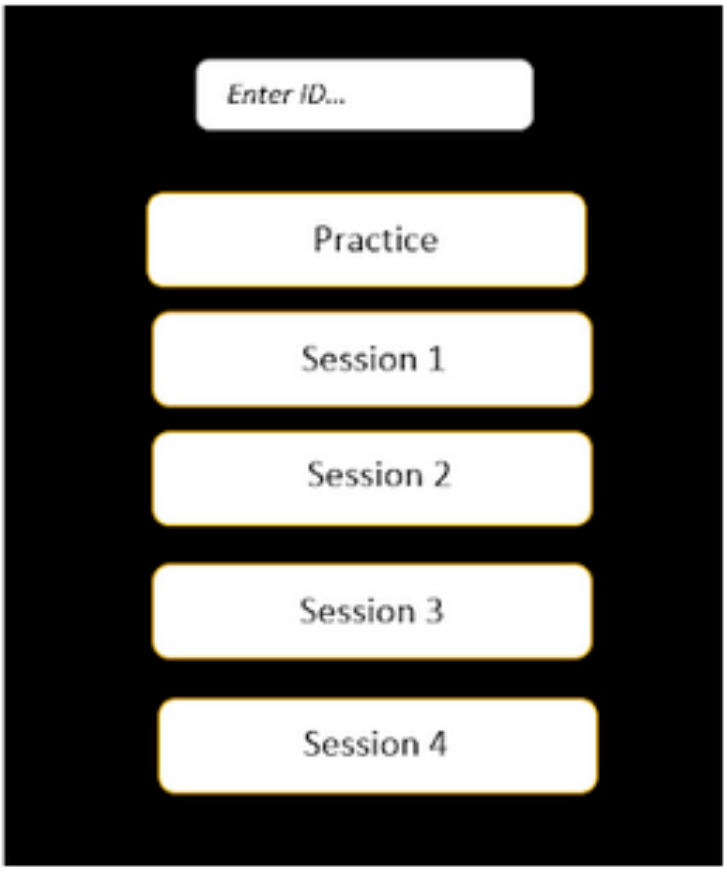


Figure 2.5 First version of the menu

Information Sheet designed for participants also had to integrate extensive instructions about the usage of the software. The update version of the Information Sheet exists in Appendices while the second version of the UI is described in section 3.7.3.

3.8.2 Version 1

The participants were expected to only take part in the encoding and recall tasks, without handling the UI, as this was the facilitator's attribution. Therefore, no transition between the UI and the encoding and recall tasks was signaled by the change in lightning, just by using a 3s countdown. Furthermore, at this

stage, the UI was offering minimum detail to the facilitator about what actions should be performed and what is the functionality achieved by each element of the UI. The facilitator was considered prepared to operate the software without guidance. The forms were not present in the UI as they were supposed to be given to the user before the experiment started.

Without the current health restrictions, the study was meant to have more participants and therefore, the set of 100 images was meant to be encoded and recalled only once. In this context, a session represents a set of 25 images. A participant was meant to encode an image and then recall it. During each session this process will repeat 25 times then the participant should take a break. The experiment would end after all 4 sessions are completed. The whole process was considered to take around 20 to 25 minutes.

3.8.3 Version 2

Overview

The functionality of each UI component is explained below. Importantly, less people were expected to participate in the study under these conditions as the participants had to own a Vive Pro Eye VR Headset. Even within these conditions, it was necessary that enough gaze patterns were collected during the experiment. Therefore, as explained in section 3.4.2, a set of 100 images was used 3 times such that the system would retrieve enough data for the classification task. Moreover, in this version, a session represents 100 images, therefore the system contains 3 sessions. Each session is divided in 10 tiers, each containing 10 images. This results in a longer experiment duration. Instead of lasting 20 minutes as it was initially designed, the experiment takes 1.5 hours. Each session still lasts 20 minutes but the participants' effort is greater and more breaks are scheduled.

The navigation of the system is done through menus that are guiding the user through the necessary actions prior to the encoding and recall tasks. Even though the Instruction Sheet is explaining every step in advance, the UI within the system contains succinct reminders of what the user should do such that they feel guided throughout the experiment. The user navigates through the UIs using the "Next" buttons.

All the buttons are made large enough such that misclicking is improbable. Hovering over the buttons with the VR controllers changes the colour intensity of each button in order to highlight the button that is pointed at. All buttons are pressed using the controller's trigger button. Every center menu is

Figure Image depicting the First Menu.

placed at 1m above the ground and 0.75m right in front of the user. The text is large and clear for these distances and the colour used is contrasting with the background such that it is not difficult for users to read the text. These specifications are consistently followed for each of the menus described below such that the user learns easily how to use the system.

ID Validation Menu

Every time the system starts, the user is asked to introduce his/her participant number. The participant number consists in the last 4 digits on the library card or, in the case of non student participants, a number was given to each user separately prior to the experiment. The validity of the ID is checked and the user can proceed only if the ID introduced contains 4 digits.

This menu shows up every time the experiment starts in order to distinguish between a new and a returning participant performing the study. Here it is decided if the ID was used before so it accesses the correct files, or the appropriate files are created for a new user. The necessary files for each user and the checking process are described in chapter 4. The keyboard used in the project represents a free asset called “VRKeys” released by The Campfire Union and available from Unity Asset Store. A few modifications were made at the source code such that it accepts only 4 digit inputs. They will be explained in chapter 4.

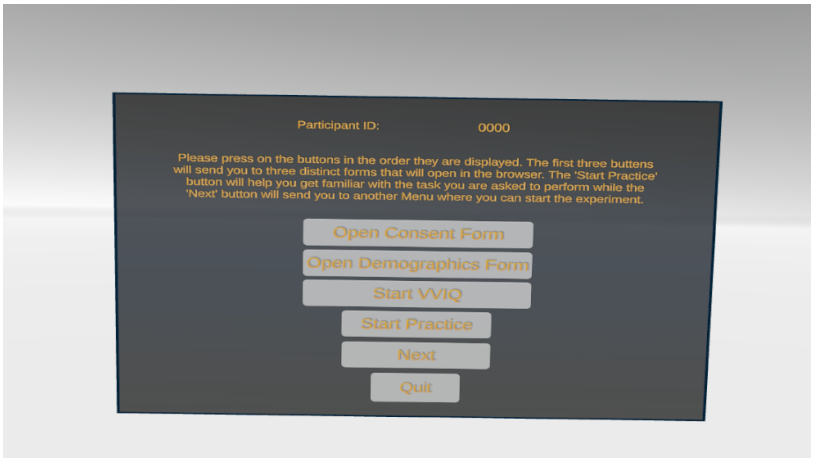


Figure 2.7 Image depicting the Session Menu.

Preparation Menu

The main purpose of this stage is to familiarise the user with the task of visual imagery and to allow the researcher to check if the eye trackers retrieve the information correctly and make sure the participant has visual imagery. At this stage, the user is asked to fill the consent and the demographics forms before they can proceed further as seen in figure 2.7.

Even though the visual imagery process was described in the Information Sheet presented to all users beforehand, completing a visual imagery task before the experiment starts is important in order to allow them to familiarise with what visual imagery means. Importantly, not all people are able to visually imagine. For some participants it might not be clear if they have the capacity to visually imagine scenes in their mind. The VVIQ represents a good indication for the user, as well as for the researcher, of the visual imagery capabilities of a participant.

Within the Preparation Menu, the user should attend a practice session for the experiment. During the practice, the user experiences the exact sequence of events present during the real experiment. This is meant to accommodate the user with the encoding and recall tasks. The practice consists of 5 images. An image is displayed for 5 seconds, during this time the user is asked to closely analyse it. After the image disappears, the reference frame remains available and the user is asked to form a mental image of the previously seen simulus while looking inside the reference frame. The process is repeated 5 times, for each stimulus. The images used during this trial are not used during the experiment

Figure 2.8 Image depicting the Session Menu.

as the users are already familiarised with them. The user can form an idea about the types of images they will see during the study.

Session Menu

The session menu allows users to start the experiment. At this point, they are considered to be knowledgeable in the encoding and recall tasks and all the questions they had so far are answered. The menu displays 3 sessions of 100 images as seen in fig 2.8. The users are meant to encode and recall the 100 images 3 times, each time in a different session. The session number they are currently completing is displayed as well as their progress (i.e. number of images encoded and recalled out of 100). Each session is divided in 10 tiers to allow frequent breaks therefore, a session is completed after all 10 tiers are completed. A new session can begin only after the previous one is completed. This was enforced such that all images are displayed at least once. The images are displayed in random order so the sequence of images will most likely differ between sessions in order to reduce any biases.

The last form a user is asked to complete is the Spatial Memory Form. This form is last presented as participants should not concentrate on figuring out the purposes of this project. Showing them the Spatial Memory Form might affect their natural behaviour during the project.

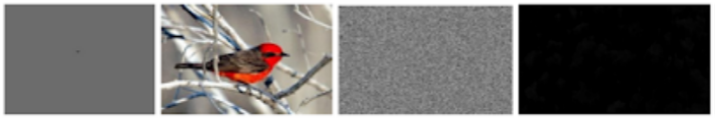


Figure 2.9 Depicts the sequence of events the system performs during the encoding and recall tasks as perceived by the user.

3.9 Sequence of events during encoding and recall tasks

3.9.1 Events

This section will describe the necessary sequence of events that have to be performed by the system in order to implement the looking at nothing paradigm correctly. The fixation cross is displayed through every event such that the user accommodates with it and its saliency effect during recall is reduced.

The user needs to have his/her attention redirected towards the area where images are displayed. This requirement is achieved using a fixation cross displayed in the middle of the field of view, coinciding with the center of the stimulus that will be displayed. The fixation cross is visible for 1s.

Immediately after the fixation cross disappears, a random stimulus is displayed for 5s and the user performs the encoding task. During this time, the eye data from each participant is stored in a .csv file.

A dynamic noise mask is displayed for 0.5s after the stimuli such that any residual information on the retina is wiped away. This is an important step as the environment is completely dark and the displayed images are the only lit object in the user's field of view leading to possible information leftovers on the user's retina. The displaying time of the noise mask is short enough such that the users will not struggle trying to recall the stimulus.

For the next 8s the only the reference frame remains displayed and the user recalls the previously seen stimulus. During this time, the eye data from each participant is stored in a .csv file.

The sequence described above is represented in Figure 2.9 and it is repeated for every displayed image.

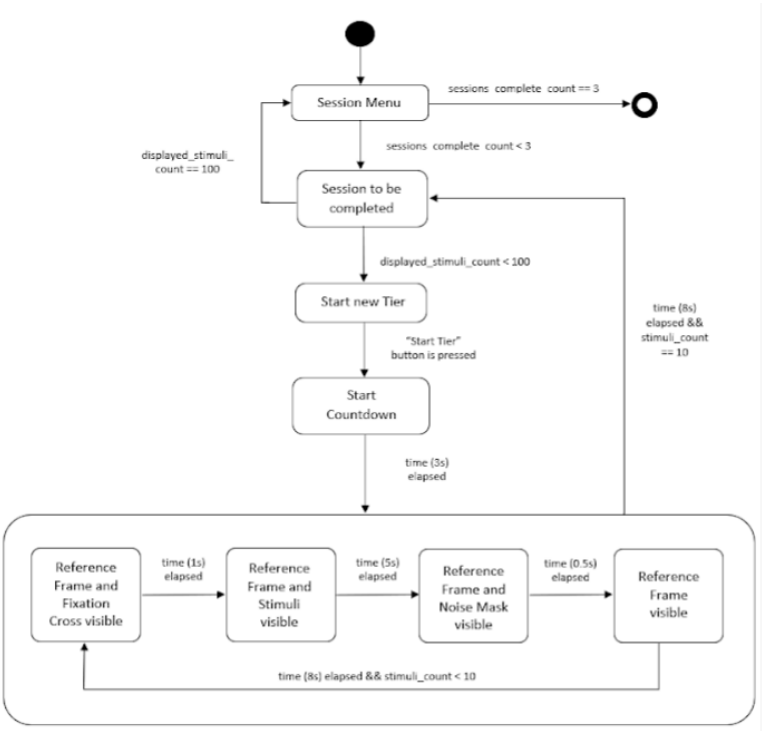


Figure 2.10 State machine diagram of the encoding and recall tasks.

3.9.2 Transitions between events

Figure 2.10 represents the state machine diagram of the encoding and recall tasks and this section will explain the transitions between each session.

Start of Experiment

A session is completed after all 100 images are viewed once. Each session is split in 10 tiers such that the user can take breaks and recover his/her focus. When a “Start tier” button is pressed the Countdown screen is displayed. Afterwords, the looking at nothing paradigm is used for the encoding and recall tasks. After the current tier is finished, the number of images viewed in the current section is checked again. If it is smaller than 100, a new tier for the current session is available. If the number is equal to 100, then the system makes a new session available for the user as long as the number of completed sessions is smaller than 3. If the number of completed sessions is equal to 3, then the experiment is over

and the user can not start the experiment again.

Loop for each session

A session is completed after all 100 images are viewed once. Each session is split in 10 tiers such that the user can take breaks and recover his/her focus. When a "Start tier" button is pressed the Countdown screen is displayed. Afterwords, the looking at nothing paradigm is used for the encoding and recall tasks. After the current tier is finished, the number of images viewed in the current section is checked again. If it is smaller than 100, a new tier for the current session is available. If the number is equal to 100, then the system makes a new session available for the user as long as the number of completed sessions is smaller than 3. If the number of completed sessions is equal to 3, then the experiment is over and the user can not start the experiment again.

Looking at nothing paradigm

As explained in section 3.8.1, the frame is visible throughout the whole tier. At the beginning the fixation cross is displayed until the time associated with it elapses. Similarly, the randomly picked stimulus is displayed for 5s, followed by the dynamic noise mask for 0.5s. Lastly, after the noise mask's time elapses, the reference frame remains the only object displayed for 8s. This sequence is repeated if the number of displayed stimuli is smaller than 10, otherwise the system goes back to the "Session to be completed" state and checks the number of total images displayed within the current session.

3.10 Data collection

3.10.1 Organisation

The eye movement data will be retrieved during the Practice task and during each tier. The data is stored in a .csv file since this type is convenient for the type of data retrieved and the post-processing stage. The .csv file name is constructed using the participant ID, the type of task and the name of the stimulus used during the respective task (¡ParticipantID¿_¡TaskType¿_¡StimulusId¿). Therefore, the name of the file containing encoding gaze data for stimulus i205847859 from participant 0000 would be 0000_Encoding_i205847859.

Even though the eye tracking is active the entire time once a tier starts,

only the gaze registered during the encoding and recall are stored in their respective files. While the fixation cross and noise mask are displayed, the gaze patterns gathered are stored in a separate file with the TaskType equal to maskAndCross. This approach has testing purposes, no information is lost and the researcher can check if the timing of the software is accurate. Moreover, this format was adopted to make sure that only the gaze patterns generated during recall or encoding are analysed and used in post-processing.

Alternatively for the Practice task, the names of the stimuli are set to 1, 2, 3, 4, 5, so a name file would be 0000_Encoding_1. The data are retrieved during the Practice task to allow the researcher to test if the participant is following the instructions and if their eye data during recall task are extensive.

All the files generated by one user are stored in a directory named as the Participant ID. These directories stored in a persistent data location on the computer used for performing the study. All the directories and .csv files generated through the experiment are created at runtime. After each participant finishes the experiment, the directory containing all his/her data are sent via email to the project owner. This adds to the deployability of the whole system such that participants are only asked to perform the encoding and recall task and complete a few forms.

The order in which the stimuli are displayed is saved in a .txt file. These files can be used to check any potential biases that can arise during the experiment due to the order in which the images are presented. These files also have other additional functionality, the software uses them to check which stimuli were already displayed and the next random image chosen was not in fact shown before.

The necessary available storage for one participant is 60MB.

3.10.2 Raw data

During the study, the gaze patterns are retrieved at a frequency of 100 entry points per second. The low frequency is imposed by the eye tracker and the Unity engine in the VR mode. The frequency number can vary at a small percentage as the game engine used for creating the study does not have a fixed refresh rate depending on what other processes are run simultaneously. Therefore, during every encoding task 500 entry points are generated and stored while during the recall task 800 entry points are generated and stored.

For every frame, the gaze data will be stored in the csv file containing the following columns:

- Participant ID: The number associated with the logged in user. [int]
- Date (format YYYY/MM/DD): The date when the experiment was performed. [string]
- Time (format HH:MM:SS.sssss): The time when the gaze patterns were generated. [string]
- Milliseconds: Time in milliseconds since the start of experiment. [int]
- GazeIsValid: Value that represents the validity of tobii eye data. [bool]
- x, y, z coordinates for the origin of gaze. [float]
- x, y, z coordinates for the direction of gaze. [float]
- Blink: Value that tracks if the eye is open or closed. [bool]
- LookingAtStimuli: Value that represents the gaze being directed towards the area where the stimuli are displayed. [bool]
- Name of stimulus displayed [string]
- Current session number [string]
- Current task type [string]

3.11 Customisation

The customizability of the software produced for this project is important as it allows other researchers implement the looking at nothing paradigm without great effort. The customizability is concerned with variables and functionality that can be changed depending on the needs and hypothesis of each study.

The stimuli dataset can vary greatly according to the needs of each study. This project used indoor and outdoor scenes but other researchers might study the involuntary eye movements made during the recall of human faces or other less complex stimuli such as inanimate objects. Allowing future users to change the dataset used during the encoding and recall offers wider usability.

So far, previous studies concerned with the eye movements made during recall did not use similar encoding or recall times. Different elapsed times for the encoding and recall tasks can give various insights about the cognitive processes performed by participants to the study. Therefore, the time allowed for encoding

end recall should be decided by each researcher depending on the needs of their study.

Depending on the dataset and the time of recall and encoding tasks used, different frequency of breaks might need to be scheduled, therefore, the tier sizes should be customisable.

Depending on the number of participants the number of sessions can vary. This project used 3 sessions to cope with the small number of participants that were able to attend the experiment considering the current health restrictions. For future studies, the number of participants might not be a problem and the number of sessions might be reduced. On the other hand, this software can also be used to generate large gaze datasets for more in depth analyses, therefore, the number of sessions might need to be increased.

Other customisable options represent the distance between the user and the stimuli and the size of the stimuli. By changing the stimuli size and also the stimuli dataset, the reference frame should also change according to the image being displayed. Therefore, the reference frame must dynamically change its shape depending on the size of each stimulus even if the data set chosen by a future researcher contains stimulus of various sizes. The system should also provide an option where the reference frame can be made visible or not during the experiment.

Since the experiment can be carried out remotely, the data collection also needs to be automated. Therefore the recipient e-mail address should be customisable.

The default values for all these options are set as the ones mentioned in this study.

Chapter 4

Implementation

4.1 Overview

In order to achieve the design and functionality explained in Chapter 3, the following Softwares and APIs were used:

Unity represents the game engine used to develop the software described in this project. The version used: 2019.4.18f1.
C is the programming language used to add additional functionality to the Unity components of the software.
SteamVR API is the tool used for adding VR support to the Unity environment.
HTC Vive Pro Eye is the VR headset used for testing and during the experiment. The software is designed for this piece of hardware.
TobiiXR API SRanipal Eye API were used in order to detect and retrieve the eye data from users.

4.2 System Logic

The Unity project is divided in 3 scenes, each being concerned with different functionalities of the software. The game logic assigned to each Unity scene will be described below.

1. Menu

   This scene contains all the UI elements of the software and it is the only

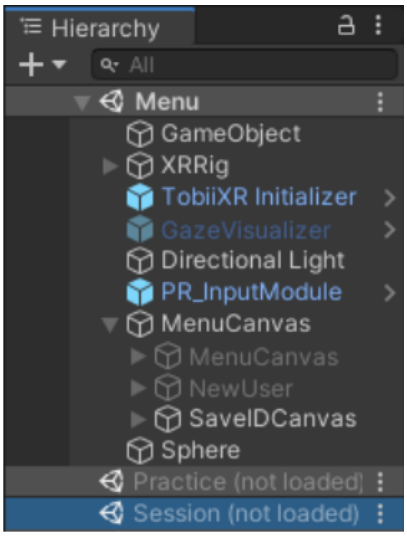


Figure 2.11 System Logic

visible scene when the system starts running. All the sub-menus (ID Validation Menu, Preparation Menu and Scene Menu) that allow the navigation of the software are located in this layer.

As this is the first scene to be loaded when the program starts, this layer needs to initialise the Virtual Reality and Gaze Tracking dedicated objects such as camera, controllers and TobiiXR Initialiser. Even though these objects are initialised when the software starts running, while this scene is initially loaded, they are persistent objects that do not get destroyed once the scene is unloaded and they stay active until the software stops running. This layer recognises new or returning participants and it also allows the next two Scenes to be loaded (and therefore be visible) by triggering the right actions (i.e pressing the "Start Practice" button or the "Start Tier for Session x" button).

2. Practice

This scene contains all the necessary elements for initiating the preparatory Practice task. Even though similar to the experiment, this layer encompases more basic functionality therefore it is separated from the Experiment scene.

3. Experiment

This scene controls the system logic dedicated to the looking at nothing paradigm, including timing of tasks, changing the lightning conditions, storing of eye data, generating all the files necessary for retrieving and storing data,

sending data to the project owner etc.

4.3 Image Dataset

The stimuli dataset has a size of 30 MB. Therefore, the stimuli have to be stored locally, on the computer used to run the software. The location used for storing the stimuli directory is set as the "Desktop" directory in Windows. The stimuli dataset can be easily changed since it is stored locally and not embedded in the software.

Each image is randomly retrieved from the stimuli data path at runtime and displayed during the encoding phase on a Canvas object of size 0.28x0.28 m. Even though the canvas is square, the image will have its aspect ratio preserved such that the largest dimension fits the canvas. In this case, the largest dimension will always be the width since all images are landscapes. Since the resolution of each image is 1024x768, the aspect ratio of each image is 4:3 and therefore, the image size on the screen will be 0.28x0.21 m.

For each session, an empty .txt file is created with a title respecting the format ¡ParticipantID¿_¡SessionNumber¿.txt. This file is used for storing the data path of the images already displayed for a particular participant during a particular session. Each time a random image is chosen within a session, the appropriate .txt file is checked line by line in order to determine if the image was previously displayed. If the file does not contain the image path, the picture is displayed for the user to encode it and the name of the image is added on a new line in the file. If the file contains the image path, a new random image is picked and the process repeats. A different file is created for each session such that all the images are displayed only once per session.

4.4 User Interface

When the system starts running, the user is asked to input his/her participant ID. This is enforced as the system needs to decide if the current user is a new participant or a returning user in order to access or create the necessary files for storing data.

This functionality is achieved using the SaveIDCanvas object and the VRKeys asset designed by The Campfire Union and available from Unity Asset Store. The user enters the ID using the VR keyboard and a controller.

The asset was designed such that the controller is used as a drum

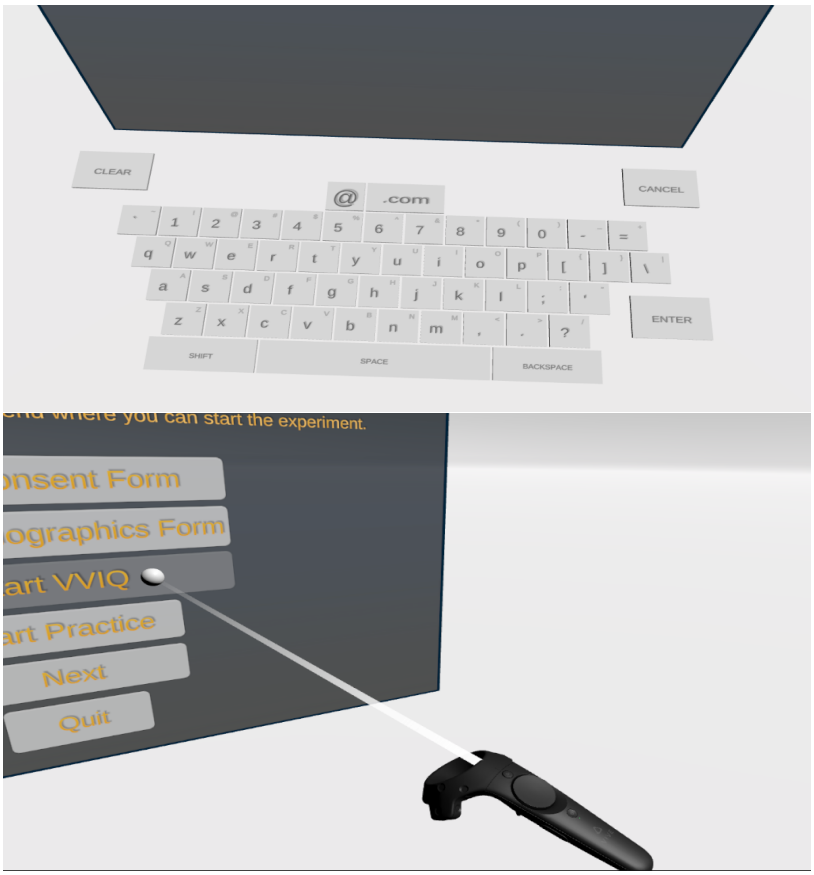


Figure 2.11 Keyboard and controller during the save id menu

stick pushing on the keys. Each key is a child object of the keyboard and they are pressed using the velocity at which the controller is pushed down above a key. After the ID is typed, the 'Enter' keyboard needs to be pressed such that the input value can be checked. The value introduced by the user needs to be a string of length 4 containing only digits. The demoScene.cs class takes the input from the VR keyboard and checks if it follows the above mentioned constrictions. After checking the validity of the input in the HandleSubmit method, the system displays one of the two messages shown in figure 2.13 using the SubmitID method.

If the user introduces a valid ID, then the Next button appears on the screen, allowing the user to proceed further, otherwise the user Is asked to introduce a valid ID. The demoScene.cs takes the valid participant ID and stores it in a public static string variable called participantID. This variable is used

Figure 2.13 Confirmation and error messages

later on in other classes.

If the user presses the Next button, the software checks if a previous user with such an ID logged in before. All the data gathered by the software is stored in a designated persistent data location where every new user gets assigned a directory named after their ID. This is where all the data generated by a user is stored. If the name of an existing folder matches the input of a user, then the system will check their progress to determine what sessions and tiers have to be made available to the returning user. The system will use the address of the existing folder to save the data generated during the experiment. If the returning user finished all 3 sessions, no tiers will be made available for them. If the input does not coincide with any existing directory, the user is considered to be a new participant and a new directory is created. The user is then redirected to the Preparation Menu.

The NewUser.cs class is used to display the participant ID at the top of

the Preparation menu. The user is then redirected to the Consent, Demographics and VVIQ forms by pressing the appropriate buttons. The Preparation Menu allows users to start the practice for the experiment. By pressing the Start Practice button, the user is redirected to the Practice scene and the trial starts. The Next button redirects the user to the Session Menu.

Each session progress is checked before the Session Menu is presented to the user. The GetProgres.cs class accesses all 3 .txt files and reads the number of lines present in each file. As each image path is written on a different line, the number of lines represents the number of images already displayed. Therefore, the progress in each session is shown to the user in the format ¡number_of_lines¿/100, as it can be seen in figure 2.9.

In order to decide which session should be completed by the user, the CheckProgress.cs class is used. Each Session has a Start tier for session ¡number¿ dedicated. At the start, the class disables the tier buttons for session 2 and 3 leaving just the session 1 tier button active. Then, the number of lines in ¡ParticipantID¿_S1.txt is checked. If the number is smaller than 100, the class leaves the 2 buttons for session 2 and 3 disabeled and the button for session 1 enabled, allowing the user to continue the experiment for session 1. If the number of lines in ¡ParticipantID¿_S1.txt is equal to 100, then the session is considered completed and the button for starting a new tier in session 1 is disabled while the corresponding button for session 2 is activated. The same process is repeated for session 3. In this way, the participant is able to perform the experiment on only one session at the time. When session 3 is completed, all buttons are disabeled. This functionality can be seen in figure 2.8.

As explained in section 3.10, one of the columns of each .csv file represents the season number. In order to set the value of this variable, the SetSessionNo.cs class is used. Each session tier has a listener attached to it. When the button is pressed a public variable named sessionNo is set to either S1, S2 or S3 such that the value of sessionNo can later be accessed and written in the .csv file. The listener attached to each session tier button also sends the user to the Session Scene where the encoding and recall tasks are performed for 10 images.

The DisplayID.cs class is used to display the participant ID at the top of the Session Menu and the Form 2 button redirects the user to the Spatial Memory form. As mentioned before, this form is given last to participants to maintain their understanding about the aims of this project minimal. This is also the reason why the name of the button is not self explanatory.

Figure 2.8 Session 1 completed, session 2 started.

4.5 Looking at nothing paradigm

The functionalities of the looking at nothing paradigm are implemented using the *NewBehaviourScript*.cs class and the Session scene. This scene contains 2 major objects - the Room (for the minimisation of distortions) and the *StimuliObject* (for displaying the sequence of fixation cross, stimulus, noise mask and reference frame) - each containing multiple child objects, as it can be observed in figure 2.14.

4.5.1 The Room

If the gaze is in fact directed toward the respective object then the class changes the original colour of the object to a different colour decided by the programmer. Black is used as the original colour and target highlighting colour such that the user is not distracted more by the inner working of the software. So this functionality is useful only for the project owner. This is one way in which the project owner checks how focused a user is on the encoding and recall tasks. However, no feedback is given to the user about their performance. One way in which the user can be notified in case they do not follow the instruction correctly, is to pause the recall or encoding task and remind the user so focus on the task and look towards the displaying area. Another approach is to check how much the user looked away from the display area during a tier and if the value is above a threshold the system would display a message to the user after the tier is done and remind them to concentrate on the encoding and recall tasks

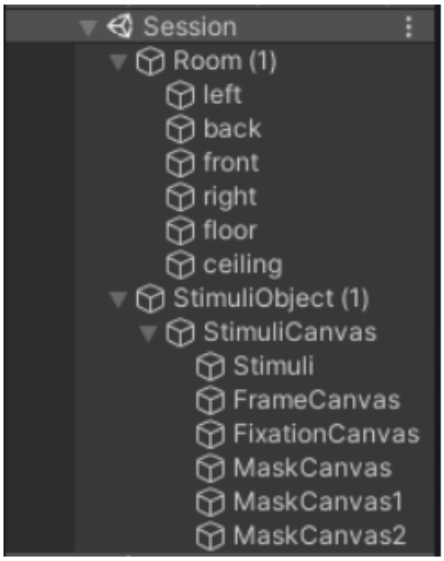


Figure 2.14 Session Scene

and look inside the designated area for both tasks. This functionality could be added in the future to maximise the user performance.

Moreover, the room is the component that implements the minimisation of distractions requirement. Every wall is pitch black, leaving the user with only one salient object - the stimulus.

The Room object is essentially a cube with black walls inside which the player is located during the encoding and recall tasks. Each wall is of size 7.5x7.5m and has attached the UIHighlightAtGaze.cs class. This enables the software to detect if a user is looking at a component of the room instead of fixating on the area where the stimuli are displayed. The UIHighlightAtGaze.cs script is a prefabricated class made available by TobiiXR API. Once the UIHighlightAtGaze component is added to an object, it detects if the gaze is directed towards the specific object the class is attached to. The bool value, hasFocus, used to check if the gaze is pointing at a wall is later used in the class that writes all the gaze data in the .csv files.

4.5.2 The *StimuliObject* and the *NewBehaviourScript.cs*

The *StimuliObject* and the *NewBehaviourScript*.cs class implement the rest of the functionalities that need to be achieved such that the looking at nothing

paradigm can be successful in facilitating the occurrence of extensive recall eye movements.

The *StimuliObject* contains 6 child objects that are used to enact the sequence of events happening during encoding and recall . Each of the 6 objects - Stimuli, Frame, *FixationCross* and 3 for the Noise Mask - displays a different sprite during the encoding and recall tasks. More complex behaviour regarding the eye data is achieved in the Update method, which will be discussed later in the data collection section.

As mentioned before, each stimulus is picked at random and retrieved from the directory at runtime. This is implemented using the *ShowRandomImage* method from *NewBehaviourScript*.cs class.

```
public IEnumerator ShowRandomImage(){

if (Directory.Exists(_imageFolderPath)) {
// Make sure the list is empty
_projectImages.Clear();
//get path for all .jpeg files in directory
string[] dirPaths = Directory.GetFiles(_imageFolderPath, "*.jpeg",
SearchOption.AllDirectories);
// if the directory is not empty
if (dirPaths.Length > 0) {
//repeats until one tier is done
     for(int a = 0; a < StimuliInOneTier; a++) {

         //pick random image
         System.Random random = new System.Random();
         int randomImageIndex = random.Next(0, dirPaths.Length);
         Debug.Log(randomImageIndex);

          //reads the path of the picked image
          string path = dirPaths[randomImageIndex];

          // transforms image to sprite
          Texture2D SpriteTexture = LoadTexture(path);
     NewSprite = Sprite.Create(SpriteTexture, new
     Rect(0, 0, SpriteTexture.width, SpriteTexture.height),
     new Vector2(0, 0));
          yield return NewSprite;
          Debug.Log("list of textures:" + path);
```

```
          //checks the .txt file that contains all the names
          //of the images already displayed
            var lines = File.ReadAllText(textFilePath);
            if (!lines.Contains(path))
            {
        // display each sprite as mentioned in requirements
            }
            //the randomly picked image was displayed before
            //so another image has to be picked
            else { a -= 1; }
        }
     SceneManager.LoadScene("Menu");
    }
}
```

As it can be observed in the above code, all the image data paths from the stimuli directory are added to an array called dirPaths. As long as the array is not empty, a new tier will start by entering the for loop. A random image is picked by choosing a random cell within the array. The image is loaded using the LoadTexture method and then transformed in a sprite using the Sprite.Create method. The .txt file that contains all the paths of the images already displayed is checked and if the file does not contain the image data path then the program will run further in order to deploy the sequence of events. If the randomly picked image was displayed before, then the loop iterator variable a needs to be decremented by one in order to find another image that was not displayed before such that all tiers contain exactly ten images. When the for loop finishes the user is redirected to the Session Menu.

In order to display the fixation cross, mask, stimuli and frame as specified in the design chapter, the following code is used inside the for loop described above, in method ShowRandomImage.

```
nameStimulus = GetStimuliName(path);
nameStimulus = nameStimulus.Substring(0, nameStimulus.Length - 5);

image.enabled = false;
type = "maskAndCross";
```

```
//display fixation cross
checkForFixationCross = true;
yield return new WaitForSeconds(FixationCrossTime);
checkForFixationCross = false;
image.sprite = NewSprite;
File.AppendAllLines(textFilePath, new[] { path });

//display stimuli
type = "Encoding";
image.enabled = true;
yield return new WaitForSeconds(EncodingTime);
image.enabled = false;

//display mask
type = "maskAndCross";
checkForMask = true;
yield return new WaitForSeconds(0.16f);
checkForMask = false;
checkForMask1 = true;
yield return new WaitForSeconds(0.16f);
checkForMask1 = false;
checkForMask2 = true;
yield return new WaitForSeconds(0.16f);
checkForMask2 = false;

//display only frame
type = "Recall";
yield return new WaitForSeconds(RecallTime);
```

The *GetStimuliName* method cuts the string containing the entire file path of the image and returns only the name of the image. This value is stored in the string variable *nameStimulus* to further be used in the .csv file name containing the eye data gathered while encoding or recalling the respective image.

Then, the string type variable is set as "*maskAndCross*" to create the .csv file name that stores the eye data while the fixation cross and mask are displayed. Then the public bool *checkForFixationCross* variable is set to true, the coroutine execution stops for 1 second and the *checkForFixationCross* is set back to false. This public bool value is then checked in the *FixationCrossResize*.cs class and while the value is equal to true the *SpriteRenderer* component of the *FixationCross* canvas is enabled such that the fixation cross is displayed

Figure The left image represents one of the 3 noise masks used, the middle image depicts the fixation cross image and the right image shows the reference frame.

on the screen for exactly one second.

Once the time for displaying the fixation cross elapses, the type variable is set to "Encoding" and the image component of the Stimuli canvas is enabled such that the stimulus is displayed for 5 seconds. Then the image component is disabeled again, the type variable is set back to "maskAndCross" and the dynamic noise mask is displayed. The process is similar to the one explained for the fixation cross. In order to make the mask appear dynamic, 3 sprites containing slightly different noise masks are fastly displayed one after the other for 0.16 seconds. When the last 0.16 seconds elapse and the public bool value *checkForMask2* is set to false, such that the third noise mask is hidden from the user, the only visible object remains the reference frame. The type variable is set to "Recall" and the coroutine execution stops for 8 seconds. The *SpriteRenderer* component of the reference frame is enabled throughout the whole *ShowRandomImage* method, making the reference frame visible during every single step described above.

The images used as Noise Mask, Fixation Cross and Frame need to be converted into sprites such that they can be rendered and displayed in the VR environment. These sprites can be seen in the figure 2.15 and they are saved and retrieved from the Asset directory within the Unity Project. The reference frame is essentially a pitch black image, such that the user can not distinguish it from the black walls of the room, with a white border around the edges.

The Fixation Cross and Noise Mask sprites need to be of the same size as the stimulus displayed during one sequence of events, while the reference frame needs to be slightly larger such that only the whilte border is visible while the fixation cross, noise mask and stimulus are displayed. In order to achieve this, the Update method from the NewBehaviourScript.cs class stores the width and height of the stimulus in two public variables:

```
width = NewSprite.bounds.size.x;
height = NewSprite.bounds.size.y;
```

The FixationCross.cs class then reads the above mentioned public values and uses them to set the size of the fixation cross sprite:

```
w = NewBehaviourScript.width;
h = NewBehaviourScript.height;

sprtRend.size = new Vector3(w / 10.0f,
h / 10.0f, 0.0f);
```

Since the unit of measurement used in Unity is meters and the stimuli used in the experiment are presented on a canvas of size 0.28x0.28m, all the dimensions of the sprites must be divided by 10 as well. In the case of the reference frame, the division is done by 9.94 such that the white border stays visible.

All sprites are positioned exactly in the same location in the environment. Since the reference frame is continuously displayed it has to be rendered last, such that all the other sprites appear on top of it. Therefore its OrderInLayer value, inside the SpriteRenderer component, is the lowest one, namely -5.

4.6 Practice

The practice task follows the same pattern as above but the images are stored in the asset directory and therefore they are more easily retrieved.

4.7 VR support

4.7.1 XRRig containing camera and controllers

SteamVR API is used in order to implement the VR mobile camera and controllers. Once the SteamVR API is added to the Unity project, the XRRig object is made persistent such that it is not destroyed between scenes and added to the first loading scene, namely the Menu scene. The XRRig object contains a prefab object named [CamaraRig] with the *SteamVR_PlayArea*.cs class attached to it,

Figure 2.16 Contains the controller material on the right and the white Skybox attached to the camera.

allowing the VR environment to be correctly rendered according to the positions of the VR base stations. The [CamaraRig] prefab contains 3 other objects, namely the 2 controllers and the VR camera. The *SteamVR_CameraHelper*.cs class is attached to the Camera object in the Menu screen such that the position of the user is recorded within the area tracked by the VR base stations. The Skybox visible in figure 2.16 is added to the camera object such that the environment appears white during the UI navigation. The Room described in the above section is covering this skybox while the user completes the encoding and recall tasks.

4.7.2 VR controller

The software uses only the right controller in order to take input form from the user. Therefore, the Controller (right) object contains two child objects: Model and PR_Pointer. The Model object has attached the SteamVR_RenderModel.cs script in order to detect the position of the controller in VR and render it. The PR_Pointer object's functionality is to determine if the controller is pointing at an object (using the Pointer.cs and the PR_InputModule.cs scripts). It contains a line renderer that by default attaches a line of 5m to the top of the controller and a small sphere attached at the end of the line such that the user can see where the line intersects with an object. The line will take the distance between the controller and the object it is pointing at and resize the default length accordingly such that the user can stay in the indicated (0,0,0) position in the environment and interact with the UI elements. Whenever the trigger of the controller is pressed while the line intersects an interactable object (such as the UI buttons) the actions attached to those objects are triggered. The registration of a user action, while pointing the controller towards a button and pressing the trigger, is implemented in the PR_InputModule object. As it can be observed in figure 2.17, the controller line renderer resizes the length of the line depending on the distance between the controller and an object.

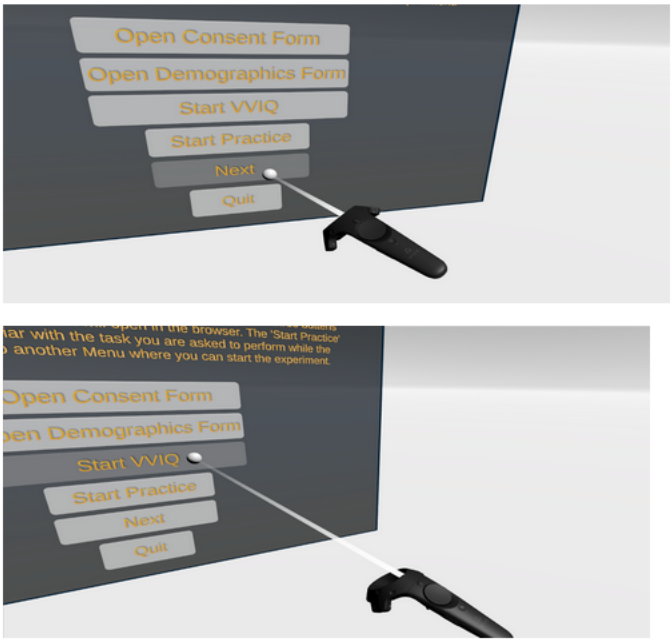


Figure 2.17 Controller line renderer resizes.

4.8 Gaze tracking support

SRanipal needs to be installed as it is the software used by the headset to detect eye movements within VR. In order to add gaze tracking support to the software, the TobiiXR and ViveSR assets are added to the Unity project. The TobiiXR_Initializer prefab designed by the TobiiXR API is added to the first scene to be loaded, namely the Menu scene, and it is set as a persistent object such that the gaze tracking is enabled throughout the Session scene and Practice scene. TobiiXR_Initializer prefab has attached a script that initialises the gaze trackers.

In order to properly detect gaze using the HTC VR headset, the list of eye trackers providers has to be ordered in the following manner - VIVE, Tobii, Nose Detection, Mouse - as it can be seen in figure 2.18. The only VR headset usable with this software is Vive Pro Eye, therefore the 'Support VIVE Pro Eye' option had to be enabled such that the SRanipalSDK runtime used by the headset is activated.

The *IGazeFocusable* interface is used to detect whether either one of the room's walls or the canvas displaying the stimuli, are gazed at. The *IGaze-*

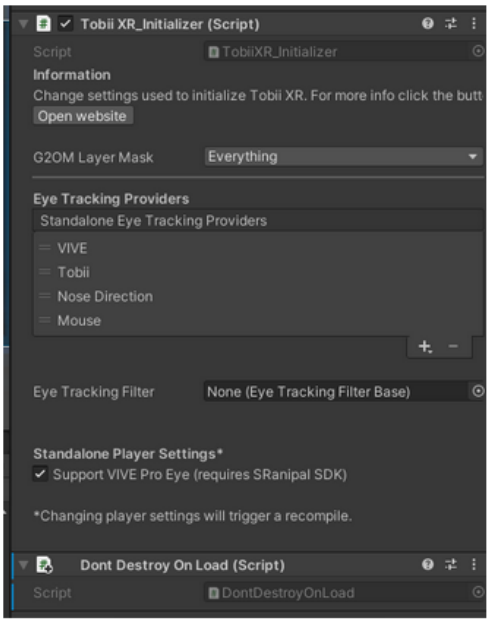


Figure 2.18 Set of options that have to be implemented such that the gaze trackers are initialised.

Focusable interface is part of the TobiiXR API and it uses collision boxes to detect if the gaze ray of a user is intersecting with the object that has a collision box attached to it.

One box collider is attached to each wall and the stimuli object, respecting the exact size of each object they are attached to. AS mentioned before, the *UIHighlightAtGaze*.cs script uses the box colliders to check if the object intersects with the gaze or not.

4.9 Data Collection

4.9.1 *NewBehaviourScript.cs*

The Update method from the NewBehaviourScript.cs script is used in order to detect the origin and direction of the gaze ray and then write this data in the appropriate .csv files. In order to generate the raw gaze data, the following code is used:

```
var eyeTrackingDataLocal = TobiiXR.GetEyeTrackingData
```

```
(TobiiXR_TrackingSpace.World);
bool isValid = eyeTrackingDataLocal.GazeRay.IsValid;
```

The information retrieved by the gaze trackers is assigned to the *eyeTrackingDataLocal* variable by using the TobiiXR method *GetEyeTrackingData* that takes as input the World tracking space. Therefore the data retrieved from the eye tracker is corresponding to the VR world coordinates. The validity of the eye data is then stored in the *isValid* bool value, to later be stored in the .csv file.

The blinks are detected using the *IsLeftEyeBlinking* and *IsRightEyeBlinking* methods and the returning values are stored in variables with corresponding names. If either of the values is true, signaling that the respective eye is closed, then the string value blink is set to "isBlink", otherwise "notBlink". The blink value is then later stored in the .csv files.

```
var isLeftEyeBlinking = eyeTrackingDataLocal.IsLeftEyeBlinking;
// The EyeBlinking bool is true when the eye is closed
var isRightEyeBlinking = eyeTrackingDataLocal.IsRightEyeBlinking;
//detect blink
if (isLeftEyeBlinking || isRightEyeBlinking) { blink = "isBlink"; }
else blink = "notBlink";
```

Then the origin and the direction of the gaze ray are retrieved using the *eyeTrackingDataLocal* variable mentioned above. The direction component of the gaze ray is represented by a normalised Vector3 variable. The 3 float points creating the vector are each assigned to a different variable and then transformed in strings such that the data can be written in the .csv file. The origin component of the gaze ray is represented by a 3D point. The same process described for the direction component is implemented for the origin as well, as it can be observed in the following code.

```
//vector for eye direction
var eyesDirection = eyeTrackingDataLocal.GazeRay.Direction;
var eyesDirectionX = eyeTrackingDataLocal.GazeRay.Direction.x;
var eyesDirectionY = eyeTrackingDataLocal.GazeRay.Direction.y;
var eyesDirectionZ = eyeTrackingDataLocal.GazeRay.Direction.z;

//3D point for eye origin
```

```
var eyesOrigin = eyeTrackingDataLocal.GazeRay.Origin;
var eyesOriginX = eyeTrackingDataLocal.GazeRay.Origin.x;
var eyesOriginY = eyeTrackingDataLocal.GazeRay.Origin.y;
var eyesOriginZ = eyeTrackingDataLocal.GazeRay.Origin.z;

//Transform float to string so the data can be written in the csv file
direction = eyesDirection.ToString("F6");
directionX = eyesDirectionX.ToString("F8");
directionY = eyesDirectionY.ToString("F8");
directionZ = eyesDirectionZ.ToString("F8");
origin = eyesOrigin.ToString("F6");
originX = eyesOriginX.ToString("F8");
originY = eyesOriginY.ToString("F8");
originZ = eyesOriginZ.ToString("F8");
```

The part of the code described so far only encompases the raw data that refers to the gaze. Therefore, this code is used to store the following values:

- x,y,z components of gaze ray origin,
- x,y,z components of gaze ray direction,
- If the user blinks
- If the gaze data gathered from the user valid according to TobiiXR API

The 2D points that represent the exact locations on the image where the user looked are calculated in a post processing stage in Python to maximise the performance of the software. Essentially geometry will be used to calculate the intersection of a vector (gaze ray formed of origin and direction) with a plane (the image). The coordinates of the vector are known and explained above, while the position of the image in the VR environment is always the same - (0, 1, 0.75). A plane needs to be defined by 3 points, therefore 3 pairs of x and y values were chosen while the z coordinate needs to remain 0.75, in order to define the plane that the image belongs to. An interaction between the plane and each gaze ray vector was determined using simple geometric equations.

Other data regarding the recall and encoding tasks needs to be stored in the .csv files.

```
date = DateTime.Now.ToString("yyyy/MM/dd");
```

```
time = DateTime.Now.ToString("HH:mm:ss.ffffff");

miliS = eyeTrackingDataLocal.Timestamp;

h = UIHighlightAtGaze.hf; //gaze ray intersect with the canvas
```

The time and date variables are calculated using the *DateTIme* class. Another variable for time in milliseconds, representing the time when a new origin and a new direction are registered by the eye trackers since the activation of the eye tracker, is stored in a variable called *miliS* by using the *eyeTracking-DataLocal* variable and the Timestamp method. The h variable represents the bool value that signals if the user is looking at the canvas or not. As mentioned before, this value is first attributed to a public variable in the *UIHighlightAtGaze* class.

Other values such as *ParticipantID*, *SessionType*, *SessionNumber* and *StimuliName* are retrieved from different classes where they are instantiated as public variables and then initialised.

In order to write all the necessary data to a .csv file, all the variables mentioned in section add and the current section are appended to a string variable named *sb* (stringbuilder). Each variable is separated by commas to fit the format of a .csv file. Then the following code is used to create the file .csv file and write the sb string on a new line.

```
filepath = string.Format("{0}/{1}/{2}_{3}_{4}.csv",
_dataFolderPath, id, id, type, nameStimulus);
writer = File.AppendText(filepath);
writer.Write(sb);
writer.Close();
sb.Clear();
```

4.9.2 Send data to researchers

The data generated by one user is then sent to the researcher using the *Email-Data*() method located in the *GetProgress3*.cs class after session 3 is completed by the logged in user.

In order to send the data of one user, the directory containing all their information and named as the participant ID number, needs to be zipped and attached to a *MailMessage* object previously created. The *MailMessage* object's subject is of the format "Participant_¡ID¿_Folder", and the address of

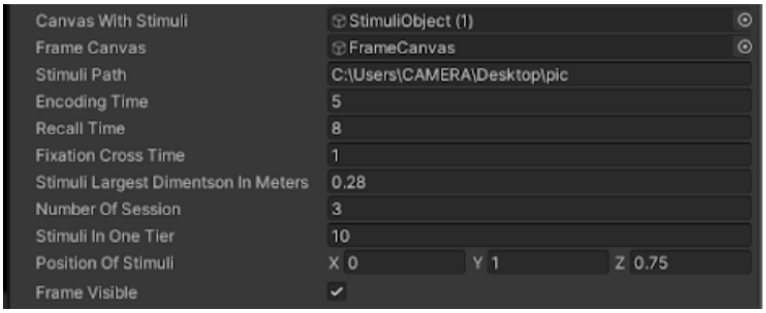


Figure Variables Representing customizable options

the recipient is set as the project owner email address. Then, an *SmtpClient* object is created using the .NET Mail API such that the *MailMessage* object can be sent to the indicated address. The credentials of the sender are hardcoded. For safety reasons these details will be encrypted in all future versions of the system.

4.9.3 Customisation

The customisation of the looking at nothing paradigm is a fundamental requirement of the software. The *NewBehaviourScript*.cs class contains 9 public variables, as it can be seen in figure 2.19, that can be changed inside the Unity engine such that the softer will comply with the needs of future researchers.

```
public string StimuliPath;
public float EncodingTime, RecallTime, FixationCrossTime,

StimuliLargestDimentsonInMeters;

public int NumberOfSession, StimuliInOneTier;
public Vector3 PositionOfStimuli;
public bool FrameVisible;
```

The *StimuliPath* is a string variable that enables the quick editing of the stimuli dataset. The *EncodingTime*, RecallTime and FixationCrossTime are 3 float variables that are used throughout the *NewBehaviourScript*.cs class, as explained before, such that the events happening during the encoding and recall tasks are timed according to the needs of the researcher.

The *StimuliLargestDimentsonInMeters* is the float variable that resizes

the *StimuliObject*. Since all the canvases displaying the sequence of events are children of this object, they will automatically be resized as well. All the child canvases are of a square shape, therefore picking the largest dimension of the stimuli and deciding how big in meters it should be is enough for the resizing to be implemented using the following code.

```
Vector3 vect = new Vector3(StimuliLargestDimentsonInMeters,

StimuliLargestDimentsonInMeters, 1);

canvasWithStimuli.transform.localScale = vect;
```

The StimuliInOneTier variable is used in the for loop used to create the necessary number of sequences in order to complete one tier within a session. The project owner needs to be aware that the number of stimuli used in the experiment needs to be divisible by the StimuliInOneTier int value, otherwise the program will not leave the for loop. Future improvements should find a solution for this edge case.

The position of the stimuli in the environment can be changed using a Vector3 object. The following code changes the position of the *StimuliObject* according to the values of the *PositionOfStimui* vector. Again, since this is the parent object, all the canvases used to display the stimuli, reference frame etc. will be relocated in the environment.

```
canvasWithStimuli.transform.position = PositionOfStimuli;
```

The FrameVisible is a bool value that enables the canvas displaying the frame when set to true and hides the frame when set to false. The following code is used to achieve this functionality.

```
if (!FrameVisible)
{
   FrameCanvas.SetActive(false);
}
```

The NumberOfSessions variable is not yet fully implemented. Future improvements should make this option fully available.

4.9.4 Creating The Installer

Since the software is meant to be used remotely by participants, the Unity project needs to be packaged for distribution. In order to do this, the Unity project containing all the scenes described above was built using the Unity engine with Windows x86_64 as the target platform and architecture. The main executable of this build together with the project directory are then inputted in Inno Setup, the software used to create the installer. The Inno Setup was chosen as it is a freely available and reliable choice for the task.

The installer enhances the deployability of the software as it is easier to use for participants than alternatives such as zip files as well as being a more memory efficient option.

Chapter 5

Testing

5.1 Overview

This chapter will explain the testing strategy developed for assessing the conformance with the requirements of the software described in chapter 4. furthermore, a high level overview of the testing plans will be described and the testing outcomes will be analysed. The testing process is broken in 2 major steps: the evaluation of the user interface and the evaluation of the encoding and recall tasks. Each of them will be described in the following sections. The system was also tested on a small number of participants due to the health restrictions and the limited number of potential participants. This section will also assess the clarity of the instructions used during the experiment.

An experiment will also be conducted in order to perform the image retrieval task. The usability of the eye data, collected using the software, as input in the classification task needs to be investigated. This evaluation will determine if the eye data gathered using the software contains enough detail given the eye tracker's limitations in order to perform well when used as input for the classification task. Furthermore, the potential future usage of eye tracking technology as a BCI technique can also be suggested after the experiment is conducted. This experiment, and therefore further testing of the software will be discussed in chapter NO.

5.2 User Interface

5.2.1 Strategy

The testing strategy regarding the user interface is to conduct preliminary tests conducted by the project owner in order to ensure that the expected functionality of the user interface is achieved. Then tests with volunteers will take place in order to verify if the UI and the instruction sheet provide enough detail about the usability of the software such that the experiment can take place remotely.

5.2.2 Planning and outcomes

The project owner conducted thorough iterative tests after every functionality was implemented. The Menu scene was the main part of the project tested at this stage. Every element of the UI such as buttons, controllers and keyboard were initially tested by the project owner. However, their evaluation is not sufficient regarding the self explanatory aspect of the software as they are aware of all the aspects related to how the software should be used. Therefore, one volunteer was asked to navigate the system according to their understanding based on the instruction sheet given to them in advance. This approach will also signal any shortcomings related to the UI functionalities or unexpected behaviour of the users that the project owner failed to account for.

The testing of most UI elements is straight forward, however for the tier buttons more in depth tests needs to be performed. It is essential that the user is able to complete only one scene at the time and that the tier buttons get activated and deactivated when necessary and that the progress displayed for each section is accurate. In order to test the functionality of the Main Menu, the time for the encoding and recall tasks was reduced within the for loop explained in chapter 4, such that the sequence of events will unfold fast and the project owner can check for unexpected behaviours. No problems were detected regarding the tier buttons or the displaying of progress. Furthermore it was necessary to investigate if the customizable aspect of the number of stimuli displayed in a tier is successfully achieved. In order to do this, multiple edge cases were tested, namely, if the number of stimuli in a tier is greater than the number of stimuli in the dataset and if the number of the stimuli in the dataset is not divisible by the number of stimuli in a tier. If such numbers are introduced in the system, an error message will be displayed, not allowing the study to continue.

5.2.3 Testing with volunteers

During the preliminary tests conducted with a volunteer the project owner was supervising the whole process in order to spot any malfunctions of the software. During the test, the following flaws were noted.

Since the start of the software, the user has already entered virtual reality. Asking participants to complete the forms in the Preparation Menu requires them to take the headset off every time they need to complete a form as they open in the browser on the computer screen, then put the headset back on to continue navigating the software. Moreover, the Consent form needs to be completed and signed before the participant enters the VR environment. The issue was noticed before the study with participants started. In order to cope with this, the Instruction Sheet asks participants to complete the 3 forms immediately after reading the instructions and before starting the software. A link to each form is provided in the instructions. An alternative solution for this problem would be to instruct participants to press the Steam Menu button on the controller and open the desktop viewer in VR. This alternative was dropped as it represented a tiring and inefficient solution that might lower the performances of participants during the experiment.

The volunteer had no problems navigating the software and the instructions were considered to be concise and clear. Moreover, all the UI elements were achieving their assigned functionality.

5.3 Encoding and recall tasks

5.3.1 Strategy

The implementation of the looking at nothing paradigm also needs to be tested. The strategy in this case is similar to the one mentioned before. The product owner investigates the correct implementation of the requirements while the tests run with volunteers assess the clasity of the instructions regarding the encoding and recall tasks. This process aims to refine the methods used during the experiment and ensure good practice when the experiment is conducted.

5.3.2 Planning and outcomes

The project owner conducted tests investigating the appropriate functionality regarding the displaying time of the stimuli during the recall task and the dis-

playing time of the frame during the encoding task. In order to assess if the behaviour of the system is as intended, a new method was implemented. It's main functionality was to start a timer when a sequence of events started to unfold and to print the total time taken for a sequence to finish as well as outputting the time since start when the stimuli canvas is enabled and the time since start when only the reference frame canvas was enabled. Furthermore, the starting and the ending timestamps recorded in the csv files also show the accurate timing of the data collected during the encoding and recall tasks. The outcome of this evaluation was that the sequence of events was perfectly timed with various values for encoding and recall time.

Moreover, the customisation of the encoding and recall times as well as the time of displaying the fixation cross needs to be tested for other edge cases. The outcomes of these tests are described in the following table.

Variable	value	Expected outcome	Outcome
Fixation Cross	0	Fixation cross does not appear	Pass
Fixation Cross Visible	1	Fixation cross appears for 1 second	Pass
Stimuli Visible	0	Image do not get displayed	Pass
Stimuli Visible	1	Image appears on screen for 1 second	Pass
Frame Visible	0	Frame do not get displayed	Pass
Frame Visible	1	Frame appears on screen for 1 second	Pass

The customisation of the stimuli position in the VR environment as well as the size of the image and the appearance of the reference frame depending on the bool value *FrameVisible* were also tested. The outcomes of these tests are successful.

5.3.3 Testing with volunteers

The second sets of tests were conducted with 2 different volunteers which did not take part in the official experiment. The first preliminary test was conducted when the first version of the UI was still in place, while the second preliminary test was conducted using the second version of the UI. The volunteers received the information sheet describing the encoding and recall tasks. They were asked to perform the practice and then encode and recall 10 images. No other information was given to the volunteers besides what was written in the instructions in order to assess how clear they are for participants. The preliminary tests were conducted in a quiet room.

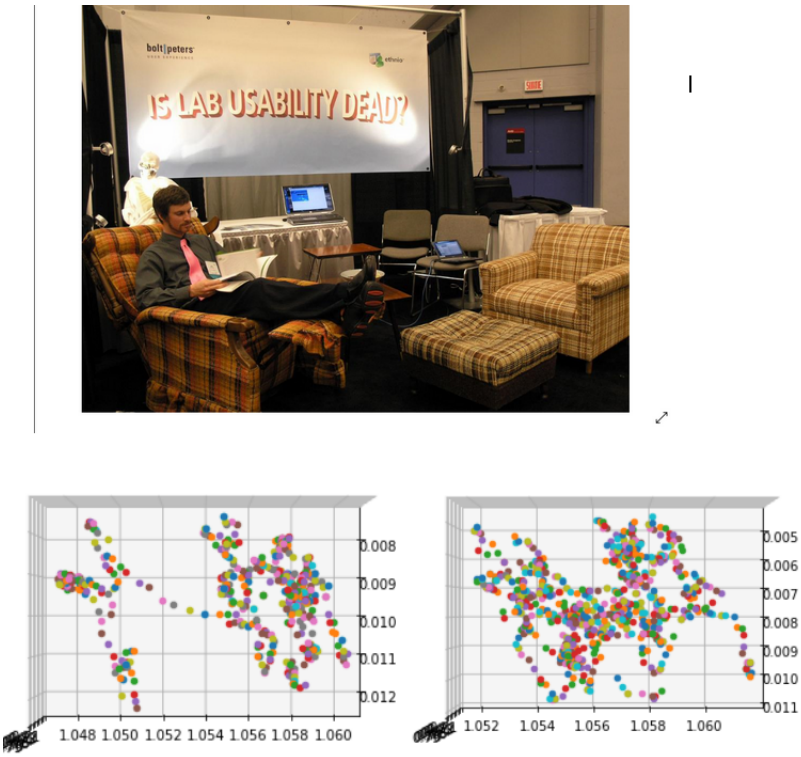


Figure 2.20 Set of data from the second volunteer after the recall instructions were revised. The top image represents the stimuli encoded, the bottom left represents encoding eye data, the bottom left represents recall eye data.

It was important to test with other people since the study owner is aware about the recall eye gaze data movements and collection, therefore, their performance during the recall task is most likely unnatural. During these tests eye gaze data was gathered from both participants and the semi-structured interviews took place after the compilation of the tasks.

The eye data was collected in order to check if the expected similarities actually emerge using the software. This was tested using python such that the data visualization could be possible. Figure 2.20 shows the sets of encoding and recall eye data for one image. Strong similarities can be observed. The full set of encoding and recall pattern visualisations can be found in the appendix.

The gaze data was gathered for 10 images, representing in fact a trial within the experiment. gathered during. As the encoding times used during the testing were the same ones as during the experiment, this set of data was

a good representation of the amount of data collected for one trial. In order for a clear idea about how much storing space was needed the size of the data gathered during this trial was multiplied by 30, resulting that around 73MB of storage space are needed for one participant. A zipped directory of this size can definitely be sent via an email so the system will be successful in sending the data to the project owner.

5.3.4 Semi-structured interviews with the volunteers

The first volunteer took part in the test while the first version of the design was in use. The volunteer mentioned they moved their eyes intentionally during the recall phase even though no such instruction was given to them, therefore the instructions were considered to be unclear. They were mentioning that the recall process should be active and that the participants should try to remember where the objects within a scene were positioned. This proved to give rise to an unexpected behaviour of the participants. Therefore, the instructions were revised to only mention that mental visual images should be formed during the recall time.

The second volunteers used the second version of the design and the revised instructions proved to be clear regarding the recall task. The information sheet used during this preliminary test and later in the experiment is located in the appendices. After creating the above-mentioned visualisations of the data, one set of encoding and recall eye movements for a stimulus was particularly good, as it can be noted in the figure 2.21. During the interview, the volunteer was asked if they remembered this image more vividly. They were not able to remember this image as being more vivid compared to the rest. But an important comment was made. They said that they did not remember a cat being in the picture and that only the cushions stood out. This observation implied that different people might encode the same image differently, but more importantly people might see new details in images in each session. This is not ideal as the classification algorithm requires gaze data with strong similarities (and therefore fixations on the same objects) to perform well. These implications will be discussed in detail in Chapter 8. Asking participants to look at the same objects for each image is not a viable option as encoding is most of the time an automatic cognitive process that is hard to control. Furthermore, participants are not provided with enough time during the encoding task to make such decisions.

As the data sample gathered during the testing phase was small compared to what is needed for all the ML classifiers, no proper classification was performed using this data. There were noticeable similarities between the two

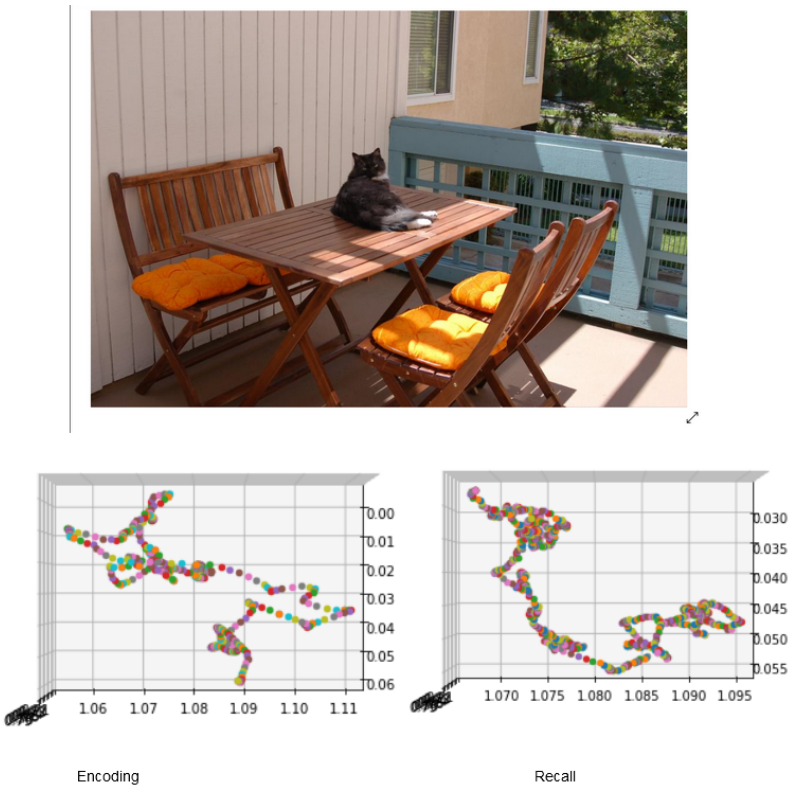


Figure 2.21 Set of data from volunteer

types of gaze data belonging to the same image. Classification using ORB key point descriptors and SSIM were tried, but the results were poor.

The testing outcomes show that all the functionalities of the software were leading towards the desired behaviours. The changes made in the instruction made the recall task clear for the second volunteer. Different people might encode the stimuli in different ways leading to different recalls as well. Instructing participants to look in the same areas of the images between sessions is not a reliable approach, as the cognitive process of encoding a stimuli is usually involuntary. Restricting the normal behaviour of users will lead to gaze datasets marked by bias. The preliminary tests also suggest the success of the software to correctly retrieve the data according to the requirements.

The quality of the eye data needs to be assessed and the following chapters will elaborate on the approach used for this task.

Chapter 6

Evaluation Study

6.1 Overview

In order to further evaluate the usability of the software described in this study, an experiment with participants was conducted in order to collect eye data for the classification task. The main priority of this step is to decide if data gathered using the software can be used to computationally discriminate, with high accuracy, images based on encoding and recall data.

6.2 Research Questions

RQ1 Can an image be computationally retrieved with high accuracy based on encoding and recall eye movements?
The aim of this experiment is to investigate the accuracy with which some classification algorithms can use the gaze data gathered by the software to computationally retrieve an image.

RQ2 Is the sequence of eye movements during recall and encoding enhancing classification accuracy when trying to computationally retrieve an image?
This experiment aims to assess if the classification pipelines using time series lead to better classification accuracies considering the limitations regarding the frequency and the 0.5°-1.1° accuracy of retrieving gaze data with the integrated eye trackers of the VR headset.

6.2.1 Variables

6.2.2 Independent Variable

There is one independent variable, which is classification type, but 6 levels to this IV will be studied (the different classifiers for both encoding and recall eye movements):

1. Logistic Regression
2. Support Vector Machine (SVM)
3. Random Forest Classification
4. K-Nearest Neighbors and Dynamic Time Warping (KNN-DTW)
5. Long short-term memory (LSTM)
6. K-Nearest Neighbors

6.2.3 Dependent Variable

The dependent variable investigated in this study is accuracy of the classifiers.

6.3 Measures

This section encompases the measures taken during the study.

Demographics: Participants were asked to complete a form asking the following details: Age, Gender, diagnosed physical conditions. The form template is located in the Appendices.

VVIQ: In order to assess the vividness of visual imagery and if the recall process is predominantly based on object imagery. The form template is located in the Appendices.

Spatial Memory: used for detecting participants with imagery processed predominantly based on spatial memory. The form template is located in the Appendices.

Semi-Structured interviews with participants after the experiment was completed in order to investigate any persistent behaviours regarding the

encoding and recall tasks among participants. The form template is located in the Appendices.

Eye Movements produced during the encoding and recall tasks.

6.4 Participants

6.4.1 Overview

The number of participants that were able to take part in this experiment was greatly reduced by the health restrictions. Only individuals who have access to the Vive Pro Eye VR headset can take part in the experiment, therefore members of the REVEAL group were targeted for this experiment as they possess the headset necessary to use the software. Further exclusion criteria were applied regarding the participants that can participate in the experiment.

6.4.2 Participant Requirements

Not have uncorrected vision abnormalities

During the encoding task it is particularly important that participants can see the stimuli clear and not struggle to process the objects contained within a stimulus image. Therefore, they should not have uncorrected vision abnormalities.

Not suffer from seizure risk, epilepsy or be pregnant

Since VR is used as the main environment of this study, participants being pregnant or suffering from seizure risks or epilepsy are excluded to minimise any risks.

Be naive about the project aim

In order to ensure that the participants behave naturally and that the eye movements performed during the recall and encoding task are accurate representations of reality, participants have to be naive about the project aim.

Be able to visually imagine scenes

As the participants were asked to perform the recall task by creating a mental visual representation of the encoded stimulus, people that are not able to perform such a task are excluded from the participants' pool. The inability to visually recall a scene in the mind is called aphantasia and participants having

such an inability can not produce eye movements related to the imagined stimuli during recall.

6.4.3 Demographics

9 participants took part in the study and a table containing their demographic data is presented below.

Participant	Age	Gender	Normal/Corrected vision
1	21	Male	glasses
2	21	Female	glasses
3	21	Male	normal
4	18	Male	normal
5	50	contact	lenses
6	21	Male	normal
7	22	contact	lenses
8	22	Female	normal
9	22	Male	glasses

6.5 Study design

Within-participants design was selected for this study as participants should take part in both recalling and encoding tasks in order to collect an appropriate set of eye data for the classification task.

6.6 Procedure

The scheduling of the experiment can be observed in table NO. People had to complete 3 sessions, each containing 100 stimuli to be encoded and recalled, dividing each session in groups/tiers of 10 images. A 1 to 2 minutes break was scheduled between each tier and a 10 minutes break was scheduled between each session such that the user can take off the VR headset and recover focus. Breaks are especially important as all participants are required to concentrate on the encoding and recall tasks in order to obtain a low noise-to-signal ratio. The completion of one session is achieved in 20 to 25 minutes and the length of

the whole experiment including breaks, forms completion and information sheet reading takes around 2.5 hours for each participant.

Event	Tier	Layout
Session 1	10 tiers	10 recall tasks, 10 encoding tasks tasks , 1-2 minutes break, 2 eye calibrations
10 min break	-	-
Session 2	10 tiers	10 recall tasks, 10 encoding tasks tasks , 1-2 minutes break, 2 eye calibrations
10 min break	-	-
Session 3	10 tiers	10 recall tasks, 10 encoding tasks tasks , 1-2 minutes break, 2 eye calibrations
10 min break	-	-

Events before experiment starts:

1. Read instructions and answer any questions participants might have. The information sheet is located in the appendices.
2. Sign consent form
3. Complete VVIQ
4. Complete Spatial Memory Form
5. Perform Practice Session

Events during completion of experiment:

1. Participants perform the encoding and recall task for each stimulus in the dataset.

2. Participants perform the encoding and recall task for each stimulus in the dataset.

Events after experiment finishes:

1. Data collected is send to the researcher via email.
2. Semi-structured interviews take place in order to find insights about the cognitive processes of the users while performing the tasks.
3. Participants are debriefed after the interviews finish such that their answers are not influenced by the new information they will receive about the experiment.

6.7 Hypotheses

The hypotheses referring to the outcomes of this study are listed below. The percentages values used are based on the accuracies obtained by Wang et al. (2020) and described in detail in sections NO and NO.

H1: Logistic Regression can achieve an accuracy higher than 90% for the classification based on encoding eye data.

H2: Logistic Regression can achieve an accuracy higher than 65% for the classification based on recall eye data.

H3: Support Vector Machine can achieve an accuracy higher than 90% for the classification based on encoding eye data.

H4: Support Vector Machine can achieve an accuracy higher than 65% for the classification based on recall eye data.

H5: Random Forest Classification can achieve an accuracy higher than 90% for the classification based on encoding eye data.

H6: Random Forest Classification can achieve an accuracy higher than 65% for the classification based on recall eye data.

H7: K-Nearest Neighbors and Dynamic Time Warping can achieve an accuracy higher than 90% for the classification based on encoding eye data.

H8: K-Nearest Neighbors and Dynamic Time Warping can achieve an accuracy higher than 65% for the classification based on recall eye data.

H9: Long short-term memory can achieve an accuracy higher than 90% for the classification based on encoding eye data.

H10: Long short-term memory can achieve an accuracy higher than 65% for the classification based on recall eye data.

6.8 Ethics

The study had to receive an ethics approval from University of Bath, therefore an EIRA1 form was completed before the experiment took place. The study received all the approvals required from the Computer Science department. The main discussion within the EIRA1 form was made around the fact that participants will not be told beforehand the aims of the studies and they do not know that the eye data during recall is collected.

During the debrief stage, all these details will be made clear to the participants in order to evaluate again their willingness to participate in the experiment as they will have all the necessary information to decide. No participant decided to withdraw from the experiment.

The EIRA1 form together with the templates used as Consent form and Debrief are available in the Appendices.

Chapter 7

Data Analysis

7.1 Overview

This section will explain the data analysis approach and the rationality behind choosing the classifiers mentioned in chapter 7. Both research questions focus on the development of a BCI with high classification performance. Each of them will be analysed separately.

7.1.1 RQs

RQ1 Can an image be computationally retrieved with higher accuracy than the current state of art based on encoding and recall eye movements?

The analyses will be conducted using encoding classification accuracies using different classifiers and recall classification accuracies using different classifiers. These accuracies will be later compared to Wang et al.'s (2020) accuracies.

RQ3 Is the sequence of eye movements during recall and encoding enhancing classification accuracy when trying to computationally retrieve an image?

The analyses will compare classifiers which utilise sequence information to the base classifiers which do not make use of the sequential information. KNN will be used as a baseline, while KNN-DTW and LSTM will be the rest of the classifiers used for this research question.

7.1.2 Data format

The raw eye data from encoding and recalling the same image are stored in 2 separate .csv files, which means that the software produces 200 .csv files that need to be analysed for each participant. In each .csv file the columns represent: ['Participant', 'Miliseconds', 'DataIsValid', 'Date', 'Time', 'DirectionX', 'DirectionY', 'DirectionZ', 'OriginX', 'OriginY', 'OriginZ', 'Blink', 'StimuliName', 'IsLookingAtCanvas', 'TaskType', 'SessionNo'].

For each time point needs to be translated in the (x,y,z) coordinates on the image. The 2D points that represent the exact locations on the image where the user looked are calculated in this post processing stage. Essentially geometry will be used to calculate the intersection of a vector (gaze ray formed of origin and direction) with a plane (the image). The coordinates of the vector are known and explained above, while the position of the image in the VR environment is always the same - (0, 1, 0.75). A plane needs to be defined by 3 points, therefore 3 pairs of x and y values were chosen while the z coordinate needs to remain 0.75, in order to define the plane that the image belongs to. An interaction between the plane and each gaze ray vector was determined using simple geometric equations. The code achieving this functionality can be found in Appendices.

The resulting 3D points are then organised within Python in two 4-dimensional matrices, one for encoding, another for recall as following:

```
all\_images\_session1 = all\_images\_session1.reshape
(num\_ppts,num\_stimuli, timepoints, datapoints)
```

Where: num_ppts = 7,num_stimuli = 100, Timepoints = 500 (encoding) or 800 (recall), datapoints = 3.

The same process is repeated for each session and then all 3 matrices are concatenated in order to create the whole dataset later used as input for the classifiers.

The main processing steps performed on the data was the elimination of all invalid eye movements based on the IsValid variable and the exclusion of blinks from the dataset. An attempt to detect the fixations within each gaze pattern was also tried. This was done using Segmentation K-Means, where k=16 as this is the average number of fixations while looking at an image. Then, the cluster center values returned by the Segmentation K-Means algorithm were used in the classifiers. This approach was unsuccessful since the saccades are

not disregarded and they impair the performance of the segmentation, leading to inaccurate cluster centers.

7.2 Research Question 1

7.2.1 Logistic regression

Logistic regression was used in order to test what is the basic accuracy achieved when encoding and recall eye data are used. Logistic Regression was chosen as their implementation is a straightforward process and the algorithms should lead to good results as long as the input data is optimal.

The sklearn Python library was used to implement the Logistic Regression classifier. The training input represents 80% of subjects (training data) while the test input is the remaining 20% of subjects (test data). These subsets will be picked randomly using cross validation in case a particular subject has distinct data. A limitation of the logistic regression algorithm is that the input array will have to be collapsed along the timepoints dimensions (timepoint * gaze data). The code used to implement this classifier and the hyperparameter tuning process (was realised using the RandomizedSearchCV library) can be found in Appendices.

7.2.2 Support Vector Machine

SVM is an algorithm that finds the best linear boundary between two classes, but it can also be used on non-linear classification problems. This is the case in this study and SVM represents an efficient alternative to classify the images as the non-linear input values are transformed in a space where the classes are linearly separable. The linear boundary of the transformed data is a non-linear boundary in the original data plotting. As in the case of Logistic regression, SVM is implemented using sklearn and the training to test ratios are 8:2. The code implementing this classifier can be found in Appendices.

7.2.3 Random Forest

This classifier is most often described as a simple but very efficient CNN. Within this algorithm, decision trees are created in order to combine the results such that a single prediction is formed. The Random Forest classifier is a promising algorithm considering the task at hand due to the minimal configuration needed

to perform well, as well as not being prone to overfitting. The following code was used in order to implement the classifier. In order to tune in the hyperparameters, GridSearchCV was used. A large number of possible parameters is imputed in this selection model and random ones are picked and tested in the classifier, the best performing ones are considered to be the optimal hyperparameter. This method was chosen as the Random Forest classifier can take a very large number of hyperparameters in order to better fit the task at hand. The code describing the hyperparameter tuning can be found in the Appendices.

```
print("Now carrying out Random Forest Classification")
from sklearn.ensemble import RandomForestClassifier
clf=RandomForestClassifier(n_estimators = 500)#r 1000
clf.fit(X_train,y_train)
y_pred=clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

7.3 Research Question 2

In the case of the second RQ, a comparison will be performed between the results achieved by simple KNN (where no sequence information is given) and a KNN which uses dynamic time warping in place of euclidean distance. KNN-DTW is a classifier that extends simple KNNs to sequences. Furthermore, the accuracies achieved using KNN on encoding and recall data will also be compared to the accuracies of LSTM.

Physiological data is usually better classified when it is considered as a sequence instead of separate timepoints. For example, KNN-DTW was successfully implemented when trying to classify muscle signals emerging during different physical activities such as walking or running (Anguila et al. (2012)). In the case of this study, sequential data might also prove to be particularly important as people tend to fixate on the most salient object first, and then proceed to observe secondary details in images. LSTM with a similar number of layers as the CNN used in Wang et al. (2020) can be used instead of such deep learning approaches. LSTM were used in the past to classify physiological metrics such as brain signals (Tirrupatur et al. (2018)). Therefore, treating the gaze data as sequential and using algorithms that are specialised in classifying data based on sequential features might lead to better accuracies during the image retrieval task.

In order to implement the above mentioned classifiers, the sequentia and keras libraries were used. The code can be found in the Appendices.

Chapter 8

Results

8.1 Classification

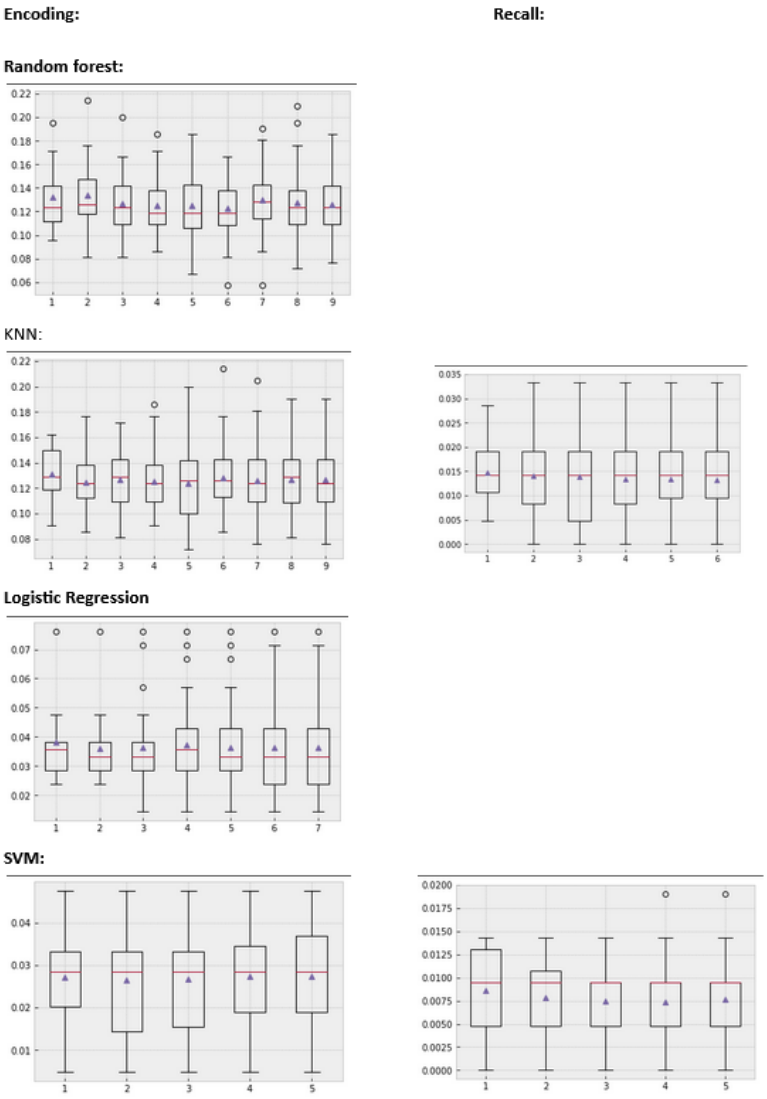
A k-fold cross validation test on each classifier, using both encoding and recall conditions, was performed in order to determine the mean accuracy across the whole dataset. The following are the results of this analysis, where little variation in the accuracies obtained depending on the dataset used for training and testing can be observed.

The k-fold cross validation test was not performed on LSTM and KNN-DTW as both classifiers take over two hours to complete one run.

After averaging across all mean values obtained in the k-fold cross validation test, the following accuracies are considered to be representative for each classifier used:

Classifier	Encoding	Recall
LSTM	0.0119	0.0109
KNN-DTW	0.0175	0.0098
Random Forest	0.1333333333	0.0261904761904
Logistic Regression	0.058333333	0.0071428571428
SVM	0.02809523809523808	0.014285714285714285
KNN	0.02142857142857143	0.014285714285714285

The best performing algorithm for both encoding and recall is the Random Forest classifier. This was the algorithm that was tuned in the most during



the pre-processing stage and it generally represents an efficient classification method for similar problem areas.

The lowest performance is allocated to Logistic Regression. Even though significant improvements in accuracy were observed after the hyperparameter tuning, this algorithm performs best for linear problems, not the case in this

study.

As a general trend, it can be observed that classification performed on encoding gaze data achieves overall higher accuracy values. This is not a surprising finding as the encoding data is in fact expected to lead to higher accuracies than recall data. The fixations performed by users during encoding perfectly overlap the locations of objects in images, and therefore the encoding datasets of different users have more similarities.

8.2 Semi-structured interview findings

- Participants mentioned looking at different details every time an image they recognised was displayed in sessions 2 and 3.
- Participants thought that not exactly the same set of images was shown in each session.
- Participants reported fatigue after a session ended.
- Participants reported looking mostly inside the reference frame during the recall task.

Chapter 9

Discussion

9.1 Overview

This section will cover two main topics: the potential reasons leading to the results described in chapter 8, and a critical analysis of the software described within this project.

9.2 Discussion of Results

The experiment described in chapter 6 was conducted in order to assess how successful the image retrieval is when the gaze data collected using the software is used as input for the classification task. All the hypotheses are rejected, as the achieved accuracies are way below the threshold values imposed by the state of the art through Wang et al. (2020).

The set of algorithms used to perform the classification tasks consists of Logistic Regression, SVM, Random Forest, KNN, KNN-DTW and LSTM. Logistic Regression and SVM were chosen as baseline classifiers for the first research question, as their implementation is a straightforward process and the algorithms should lead to good results as long as the input data is optimal. The Random Forest classifier is a promising algorithm considering the task at hand due to the minimal configuration needed to perform well, as well as not being prone to overfitting. KNN was used as a baseline for the second research question regarding the classification accuracy when the input is the gaze data represented as a sequence. KNN-DTW and LSTM were used for the second research question as well because they represent efficient algorithms that work

well with physiological sequence data, among others.

The encoding accuracy values are higher than recall accuracy values for every single algorithm. As the stimuli are presented to the user during the encoding phase, the fixation eye movements perfectly overlap with the objects within a scene that attract a user's attention. The similarities in the encoding gaze data are greater between different participants and therefore, the classification accuracy is higher.

The image retrieval task requires high similarities between eye movements of different participants while looking at the same image or visually imagining the same stimulus. The classifiers used as baselines should perform well when provided with encoding gaze patterns according to Wang et al. (2020). A potential cause for the low performance is the fact that the same 100 images were encoded 3 times. This approach was chosen because of the limited number of participants, but this is not an ideal scenario. The semi-structured interviews with the participants revealed that most participants believed the stimuli dataset was not identical between sessions, therefore the focus was reduced, probably towards the end of each session. Moreover, the length of the entire experiment and its fairly unengaging nature probably lead to a low signal-to-noise ratio even though frequent breaks were scheduled.

Another frequently mentioned observation by the participants is that they were seeing new details when images they recognised were displayed during the second and third session. Again, this is not favorable for the classification accuracies, as the gaze patterns need to be similar such that the algorithms can perform well especially for such a low number of participants. The approach of having multiple sessions, each containing the same stimuli dataset, could still have its benefits. A visual scene used in the looking at nothing paradigm can be described in terms of the salience effect of the objects within the respective scene. For example, if an image depicts a person in the foreground and an outdoor environment such as a street in the background, the person can be considered a first order detail within that picture, while different objects composing the street view can represent the second order details. In this example, the person is considered to belong to the first order as it is in the foreground. If such an image is encoded by a participant, the first fixations will probably be related to the depicted person. Some other fixations will also be attracted by the second order details. Considering that a street view can include numerous objects, different participants can pick different second order details to fixate on. The benefits of an approach where the stimuli dataset is encoded and recalled multiple times using multiple sessions come from the fact that the bias for second order details would be evened out. For such an approach to be successful in terms of the

classification accuracy scores, the number of participants needs to be high and more complex classification pipelines to be used. The extensiveness of the recall eye movements withins such an experiment still have to be investigated. Future studies might investigate the correlation between visual imagery and memory retrieval using such an approach.

The most important drawback in the classification approach within this project is the fact that 2D points were used as input for the classification algorithms. This is the most probable cause of the low accuracy scores of all algorithms. Wang et al. (2020) transformed all the gaze patterns in 2D histograms, a compact and efficient way of representing data. Various other representations for the gaze data could have been used such as 3D histograms where the third dimension is represented by time. The ideal representations of the features before performing the classification a is a vast topic and it probably deserves a separate project dedicated to it. Due to the time constraints, the input for the classifiers were the 2D points corresponding to the eye movements performed during encoding and recall. The results obtained in this project reinforce the idea that an extensive pre-processing stage for refining the data is necessary, an approach used in Wang et al. (2020).

Some time was dedicated to the hyperparameters' tuning, such that the algorithms would better fit the image retrieval task. Sklearn functions such as RandomizedSearchCV or nested for loops were used to tune the hyperparameters of each algoritam and slight improvements were obtained. Given more time, this process would continue, in order to further adapt the classifiers for the task of image retrieval. Exploration of other deep learning algorithms is another major topic and it probably deserves a separate project dedicated to it. Experimenting with such classification pipelines (and mapping pipelines if necessary) would also be another important step in order to properly answer the research questions of this project.

The classification based on recall eye movements suffers from the same limitations stated above and the accuracy is even lower as the recall gaze patterns still suffer from distortions.

The results obtained after conducting the evaluation described in chapter 6 are not conclusive about the usability of the gaze data, collected using the software, in a classification task aiming to computationally retrieve an image. As explained above, Wang et al. (2020) had an extensive stage of pre-processing applied to the data before the classification task started. Until a similar approach is implemented for this software, no definite conclusions can be stated about the software not being successful at collecting eye data that are suitable for high accuracies in the classification task. Apart from the data formatting,

another limitation of the current approach is the low number of participants. In order to perform the classification, enough data needs to be collected using the software. Using the same dataset in multiple sessions is not an ideal approach. Since the experiment ended up taking 3 times longer than originally planned (as 3 sessions were used), participants might experience loss of concentration, leading to eye data marked by noise. Event though these are not ideal conditions, it was the only feasible option that could be found considering the current health restrictions. Therefore, even though the classification task was unsuccessful, given the limitations explained in this current section, more evaluations need to be conducted in order to assess how successful the image retrieval is when the gaze data collected using the software is used as input for the classification task.

9.3 Discussion of Software

The Evaluation Study investigated if the data gathered can be used to classify with high accuracy encoded and recalled images. This experimental design was chosen since it focuses the most on developing a BCI with high classification performance. Its poor results are not irrevocably tied to a system failure to retrieve qualitative data, but mostly due to time restrictions.

However, this project can follow multiple directions. An important aspect about the software that should be evaluated is the efficiency of the reference frame to reduce distortions in recall gaze patterns and guide the user through the recall process. Such an evaluation asks for a more complex experiment, where two versions of the looking at nothing paradigm are used: one displaying the reference frame throughout all sequences of events and one where the reference frame is not used during the encoding and recall tasks. Such an experiment is possible considering the customization levels of the software. The main limitation during the Covid-19 pandemic is the number of participants that could take part in the experiment. Not enough participants were available for this to be tested within this project. With a between participants design, 30 people in each group should take part in this evaluation. Even if the image dataset was scaled and less participants were used, conducting the experiment would be extremely time consuming considering the amount of data needed.

Other experiments concerned with different encoding and recall times or other types of datasets can be conducted using the software. Even though the usability of the collected eye data can not be confidently assessed at this point, the software still successfully implements the looking at nothing paradigm and other customisation options.

By conducting the experiment described in chapter 6, the UI design and implementation were further tested. None out of the 9 participants using the software had any problems arising while taking part in the study.

Furthermore, the remote data collection was successful as the project owner received all the data generated by the remote users. This can have many benefits among which it is worth mentioning the following: studies that need to collect data fast or very large amounts of data can use this software to speed up the process of data collection, as multiple users can perform the study on their own.

If the evaluation study described in chapter 6 would be free of the limitations discussed previously and the resulting accuracies are still below the thresholds set by Wang et al. (2020), then the frequency and accuracy of the eye tracker used are most probably the causes of the classifiers' low performance. The looking at nothing parading was implemented on a technology that does not have the most performant eye trackers. As mentioned before, other eye trackers used to detect the recall eye movements have frequencies of collecting eye data 6 to 10 times higher than the integrated eye trackers in Vive Pro Eye. The extremely low accuracy in the results section might also be due to the eye tracker's medium accuracy for detecting the eye movements and the error accumulation. Moreover, all eye trackers need to be calibrated once in a while. During the experiment, they were calibrated once every 10 min - but no feedback on how often the calibration should be repeated is in fact given by the VR headset embedded eye tracker SDKs, Tobii and SRanipal.

Chapter 10

Impact of COVID-19

The impact of the pandemic was mentioned throughout the document, this chapter encompases all the limitations faced by the project due to the current health restrictions.

10.1 Change in design

Important changes in the software design had to be made in order to accommodate the remote data collection and participants performing the experiment on their own in a non-lab environment.

The software has to contain a User Interface that guides participants throughout all the steps of the study and a detailed information sheet has to be created in order to guide participants through the process of setting up the VR environment, what are the functionalities of the software and how to perform the recall and encoding tasks. These changes were made in order to cope with the lack of a study facilitator during the experiment.

The data collection had to be automated and without any involvement of the user such that the experiment can be conducted remotely. Additional steps had to be performed in order to package the software and create an installer that is appropriate for distribution over the internet and easily used by participants. All these changes proved to be time consuming.

Additionally, since the conditions in which the experiment was conducted could not be controlled by the researcher (i.e. quietness of environment), the gaze data collected from participants might suffer from low signal-to-noise ratio.

10.2 Number of participants available

In order to properly test the quality of the eye data gathered using the system and therefore test if the classification accuracies can be improved compared to the current state of art, enough gaze data needs to be collected. The health restrictions strongly affected the number of participants that can take part in this experiment. After conducting a G*Power analysis, results that 30 participants are needed for this experiment. 9 participants were able to take part in this experiment but 2 of them did not move their eyes much during the recall task, therefore their gaze data had a negative impact on the classification accuracy so this data was disregarded. would need 30 participants.

The alternative was to ask the participants to view the 100 images more than one time. At least 2800 gaze patterns have to be collected for each task. As the expected number of the participants was 10, each participant had to encode and recall the image dataset 3 times in order to cope with the reduced number of participants. This practice further impacted the classification task, as people did not perform similar enough gaze patterns between sessions. Even though risky, this approach was the only viable option for the classification task to be performed efficiently. Choosing 300 different images to be encoded and recalled by such a small number of participants is not a good input option for the classifiers.

All these workarounds might lead to low accuracy values for the classification task and deeply impaired quality of the eye data.

Chapter 11

Conclusion

11.1 Contribution

This project proposes a novel BCI where widely available eye trackers are used. However, the efficiency of this approach needs to be further studied. Even though the results are not conclusive regarding the quality of the eye data collected by the software and an improvement in the classification accuracy was not accomplished, there are still some important contributions of this project that should be noted.

A comprehensive guide based on the literature on looking at nothing paradigm was created and implemented in VR. This results in a software that makes use of virtual reality and gaze tracking in order to collect the eye movements performed during encoding and recall of visual stimuli.

The customizability of this software allows future researchers to use the product as an asset in future experiments interested in investigating the eye movements emerging during mental visual imagery.

The system's deployability in remote scenarios was tested through the experiment conducted within this project. Users had no troubles when using it and the data collection is done successfully in a remote manner. Such a software can be useful in scenarios where experiments can not take place face to face or large amounts of data need to be collected.

After conducting the experiment, a dataset of 2100 encoding gaze patterns and 2100 recall gaze patterns was produced. These datasets can be used in future for training classifiers.

11.2 Future work

More analysis on the eye data (such as reporting how much of the eye-movements fall in the picture, how many blinks, statistically different fixations between recall and encoding) should be conducted in order to form a better idea about the focus levels of users while completing the tasks in VR.

The main priority in future work regarding this project is to use different forms of data representations (such as histograms) as input for the classifiers and determine with certainty if the quality of the eye data gathered from the system comply with the needs of a BCI. Other classification and mapping pipelines should also be considered. If the data will turn out to be of insufficient quality, the system should be shifted to a normal 2D computer screen and professional eye tracker kits should be supported by the platform. Once VR headsets with more performant eye trackers are developed, the system can shift back to the VR paradigm it was initially designed for.

Other future improvements can be represented by implementations of the looking at nothing paradigm for sound, 3D objects (such that depth and imagery can be studied), word based and not image based stimuli. Such functionalities can broaden the range of experiments that can be performed using the software described in this project.

Another important future improvement is to add a user interface for the customizable options of the system. Eye data visualisation options such as heatmaps can also be added to the system.

Chapter 12

Appendices

12.1 Code and installer

Unity:

https://computingservices-my.sharepoint.com/:f:/g/personal/amd77_bath_ac_uk/Ema1anan7Y5JrCj_wCKR4IwBGNJSWJD8PzbYfPW6nj3vDg?e=o315mt

Python and Software Installer:

https://computingservices-my.sharepoint.com/:f:/g/personal/amd77_bath_ac_uk/EoLAmTVfOlxLgIeVF_-S564BjhxzFifWZeDXNmP5Zn7F1g?e=ErZwGt

12.2 Forms

Demographics:

https://forms.gle/3SJUyFjdLhGL3QRg6

VVIQ

https://forms.gle/AgUwerLbx5zwaZ639

Spatial Memory

https://forms.gle/YAmuyjAxsUq4m4xk7

12.3 Raw data and Images

https://computingservices-my.sharepoint.com/:f:/g/personal/amd77_bath_ac_uk/EiMt4x4U6UlMoaB4CNL6eO8BXZASWjBFSObNXVVL07bbTQ?e=XSJs0G

12.4 Information sheet

pdfpages

PARTICIPANT INFORMATION SHEET

Prediction of perceived complex visual stimuli based on eye data during perception and visual imagery in a VR environment

Name of Researcher: Ana-Maria Dobre
Contact details of Researcher: amd77@bath.ac.uk

This information sheet forms part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. Please read this information sheet carefully and ask one of the researchers named above if you are not clear about any details of the project.

1. What is the purpose of the project?

This study will take approximately one hour and it aims to improve on a 'Nature' study where images were computationally discriminated based on eye data. We believe that their accuracy and performance can be improved using a VR environment and different classification methods. You will be asked to perform a computer-based perception and visual imagery task. More details about this will follow.

2. Who can be a participant?

All participants aged over 18 that are not pregnant and do not have any of the following conditions can take part: uncorrected vision abnormalities, lack of visual imagery, seizure risk, epilepsy, known family members with epilepsy or recurrent fainting spells.

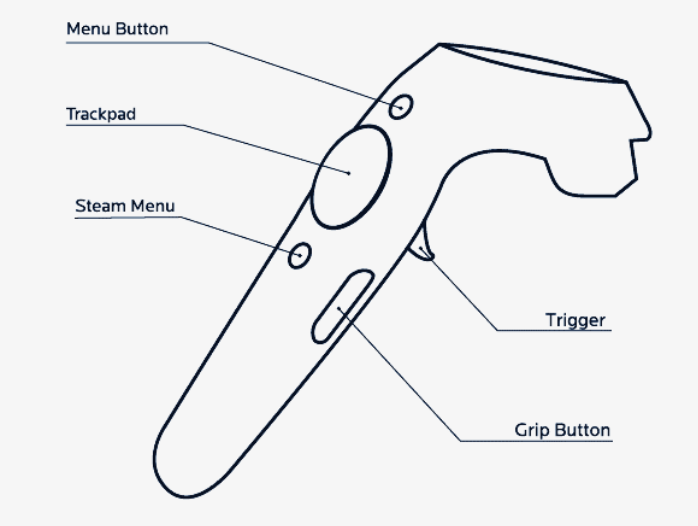
3. Do I have to take part?

It is completely up to you to decide if you would like to participate. Before you decide to take part we will describe the project and go through this information sheet with you. If you agree to take part, we will then ask you to sign a consent form. However, if at any time you decide you no longer wish to take part in this project you are free to withdraw, without giving a reason, within two weeks after taking part in the experiment.

4. What will I be asked to do?

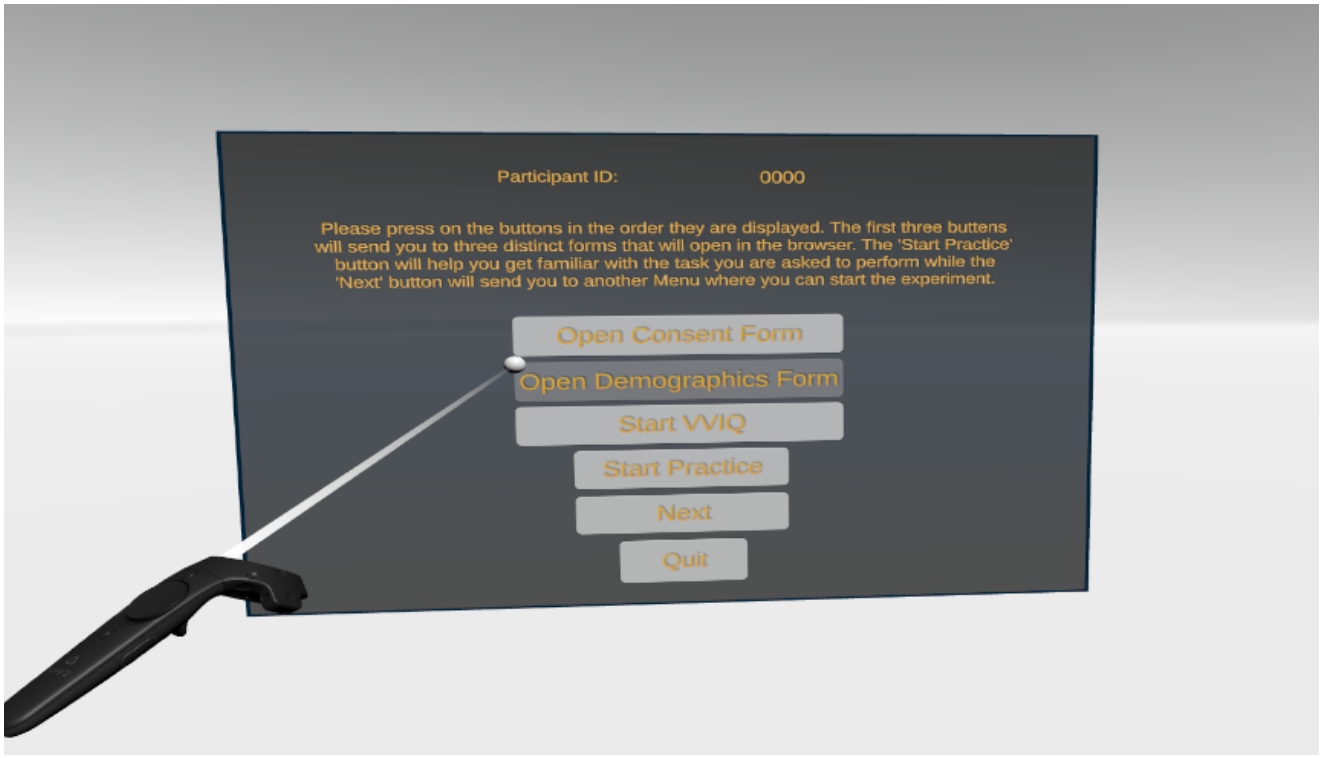
During the experiment you will be seated in a quiet room to minimise distractions. Here, you will use a Vive Pro Eye VR headset with integrated eye tracking. The experiment will take roughly one hour. You will have to complete a few forms that will open in the browser. You can visualise and complete them while being in the VR environment or by looking at a classic 2D screen without the headset, it is your choice.

You will see a green sphere on the ground, and this is where you need to stay during the whole experiment. The study is designed for you to be seated down all the time so please place the chair exactly on top of the sphere. You are advised to move your head as little as possible. You will navigate through the study using the Vive controller (importantly, only with your right hand).



When you first put the VR headset on, you will see an initial menu for a new participant. You need to introduce the last 4 digits of your library card or the participant ID assigned to you if you are not a student. You will use the VR keyboard for this. To introduce the ID, you need to push down the keys using the small sphere positioned at the end of the controller. To save the ID you need to press the 'Enter' key on the keypad. If the ID you introduced is valid, the 'Next' button will appear, and you can click it using the trigger on the controller. **In order to type the ID, you use the controller like a drum stick, it will detect the key you pressed using the velocity of your hand. In order to press on any button on the blue menus use the trigger while pointing at a button.**

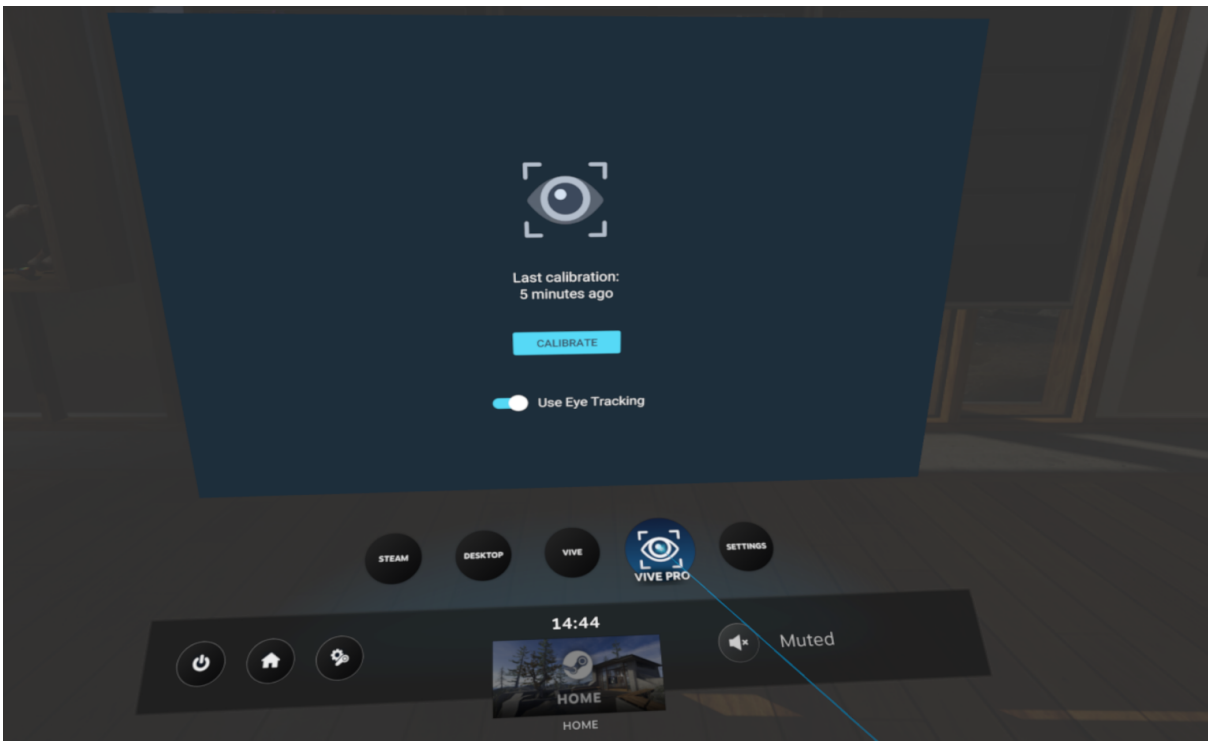
You need to go through a few introductory steps so that you will get familiar with the task you have to do. The next menu that will appear looks like this:



We ask you to click the buttons in order (from top to bottom) as they increasingly familiarise you with the task. The buttons are clicked using the trigger button on the back of the controller.

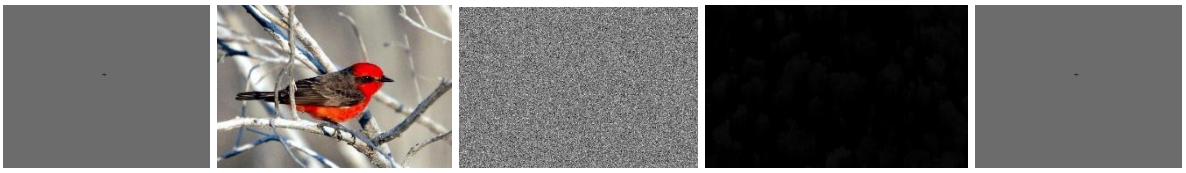
Firstly, you will be asked to open the consent form (https://forms.gle/C1GtD4TpEN2kdz2m8), read it and sign it in order to further participate in this experiment. Secondly, you should open and complete the Demographics form. The following button, “Start VVIQ”, will send you to a Vividness of Visual Imagery Questionnaire (https://forms.gle/XxZffKjzdKf7nUVk7). This will allow you to familiarise yourself with what a visual imagery task is.

After finishing the VVIQ we will ask you to press the “Steam Menu” button on the controller and start the calibration for eye tracking then go back to the starting Menu by pressing the “Steam Menu” button again. **You need to do the eye calibration before starting a new Session. A session contains 100 images and you have to complete 3 Sessions in total, therefore doing the eye calibration 3 times.**



The next button you need to press is the 'Start Practice' button. The menu will disappear, and a trial will start. This is an important step as you will see how the actual experiment will unfold. You will see the following sequence on the screen:

- A countdown for 3 seconds.
- A blank image with a fixation cross for you to direct your gaze to the centre of the field of view, it will be displayed for 1 second.
- An image depicting a natural scene will be displayed for 5 seconds and **you are asked to look at it carefully**. We refer to this as the perception task.
- A noise mask will be displayed for half a second
- Only the frame of the previously displayed image will be visible. For the following 8 seconds we want you to think about the previously seen picture and use the frame as a reference for the size of the picture. We refer to this as the visual imagery task. **You need to actively recall the image <u>with your eyes open</u>.**



1s 5s 0.5s 8s 0.5s

This sequence will be repeated 5 times during the trial, and it resembles exactly what will happen during the actual experiment, the only difference being that you will see more than 5 images. **It is especially important that you keep your focus on the image and on the <u>inside</u> of the frame during the perception and imagery task and try not to look away from them**. We scheduled multiple breaks during the experiment, but if you feel like you get distracted please pause the experiment using the Steam Menu button on the controller. This trial is just meant to familiarise you with the task and for the researchers to check if the data gets recorded correctly. If anything is unclear, you are encouraged to ask the research team for clarifications during the practice since it is particularly important that during the actual experiment you only concentrate on the visualisation tasks. **We will also**

ask you to put aside (out of the field of vision) the controller during the perception and visual imagery tasks to minimise distractions and to avoid moving or turning your head, keeping the eyes on the visual stimuli.

When the practice ends you will be redirected to the previous Menu. The next button that needs to be pressed is the “Next” button and you will be redirected to the Main Menu:

Here you can start the study. As mentioned before, it contains 3 sessions, each with the same 100 images displayed in different order. Each session is divided in blocks of 10 images, so for example you will press the button “Start a tier for Session 1” 10 times in order to complete the first session. After you finish the first session you can start the second one and then the third. Each session lasts for roughly 20 minutes but you are kindly asked to take breaks and take the headset off when you lose focus. When you start a tier, you will be redirected to the same environment used in the trial, but this time 10 images will be displayed. You need to look at a randomly chosen image and actively analyse it, then recall the previously seen image while only the frame is displayed, just as you did during the practice. The more engaged you are in the visualization and recall of the image the better. After the 10 images are visualised and recalled you will be redirected to the above menu where you can also see your progress. You can start a new session only when the previous session’s progress is 100/100. The experiment is created using blocks of 10 images so you can take frequent breaks and avoid losing focus. During the break you can take the headset off. Make the breaks as long as you need in order to recover focus.

After finishing the third session the data will be sent to the researcher via mail. Please wait and do not close the program even if it looks stuck for at least 10 minutes.

After finishing the 3 Sessions you need to complete one more set of questions by pressing the “Open Form 2” button. It will redirect you to the final form that we will ask you to complete, containing some questions about memory. After this, the study is done.

5. What are the exclusion criteria?

You must be aged over 18 and not be pregnant, nor have any of the following: uncorrected vision abnormalities, lack of visual imagery, heart conditions, seizure risk, epilepsy, known family members with epilepsy or recurrent fainting spells.

6. What are the possible disadvantages and risks of taking part?

There are no disadvantages to you taking part in the project.

7. Will my participation involve any discomfort or embarrassment?

We do not expect you to feel any discomfort or embarrassment if you take part in the project. Since this project makes use of a virtual reality headset, you may experience VR sickness. To cope with this unlikely event, we have scheduled breaks between the sessions, and you may pause or stop using the headset at any time if you feel any discomfort. You will also be exposed to a large number of natural visual stimuli. If you start losing focus you are also advised to pause the study at any time.

8. Who will have access to the information that I provide?

Only the research team will have access to information that you provide. All records will be completely anonymised.

11. Who has reviewed the project?

This project has been reviewed through the EIRA1 process within the Department of Computer Science.

12. How can I withdraw from the project?

If you wish to stop participating before completing all parts of the project you can inform one of the above identified researchers via e-mail. You can withdraw from the project at any time without providing a reason for doing so and without any repercussions.

If for any reason you wish to withdraw your data, please contact an identified researcher within two weeks of your participation. After this date it may not be possible to withdraw your data as some results may have been anonymized. Your individual results will not be identifiable in any.

13. University of Bath privacy notice

The University of Bath privacy notice can be found here: https://www.bath.ac.uk/corporate-information/university-of-bath-privacy-notice-for-research-participants/.

14. What happens if there is a problem?

If you have a concern about any aspect of the project you should ask to speak to the researchers who will do their best to answer any questions.

15. If I require further information who should I contact and how?

Thank you for expressing an interest in participating in this project. Please do not hesitate to get in touch with us if you would like some more information.

Name of Researcher: Ana-Maria Dobre
Contact details of Researcher: amd77@bath.ac.uk

12.5 Consent and Defrief

CONSENT FORM

Prediction of perceived complex visual stimuli based on eye movements during perception and visual imagery in a VR environment

Please initial box if you agree with the statement

1. I am aged over 18. ☐
2. I do not suffer from any uncorrected vision abnormalities. ☐
3. I do not have from any of the following conditions: lack of visual imagery, seizure risk, epilepsy. ☐
4. Do not have family members with epilepsy or recurrent fainting spells. ☐
5. I am not pregnant. ☐
6. I have been provided with information explaining what participation in this project involves. ☐
7. I have had an opportunity to ask questions and discuss this project. ☐
8. I have received satisfactory answers to all questions I have asked. ☐
9. I have received enough information about the project to make a decision about my participation. ☐
10. I understand that I am free to withdraw my consent to participate in the project at any time without having to give a reason for withdrawing. ☐
11. I understand that I am free to withdraw my data within two weeks of my participation. ☐
12. I understand the nature and purpose of the procedures involved in this project. These have been communicated to me on the information sheet accompanying this form. ☐
13. I understand and acknowledge that the investigation is designed to promote scientific knowledge and that the University of Bath will use the data I provide only for the purpose(s) set out in the information sheet. ☐
14. I understand the data I provide will be treated as confidential, and that on completion of the project my name or other identifying information will not be disclosed in any presentation or publication of the research. ☐

15. I understand that my consent to use the data I provide is conditional upon the University complying with its duties and obligations under the Data Protection Act. ☐

16. I understand that I am required not to disclosure any details about the project with future participants in the study before they got a chance to take part in the study. ☐

17. I hereby fully and freely consent to my participation in this project. ☐

Participant's signature: _____ Date: _____

Participant name in BLOCK Letters: _____

Researcher's signature: _____ Date: _____

Researcher name in BLOCK Letters: _____ _____

If you have any concerns or complaints related to your participation in this project please direct them to the researchers named above.

Debrief Sheet

Background:

This study aims to improve the performance of a recent research conducted by Wang et al in 2020. Numerous studies (Brandt and Stark 1997, Johansson 2006, Johansson and Johansson 2014, Laeng et al 2014, Richardson and Spivey 2000) show that during visual imagery people move their eyes involuntarily as if the stimulus they are thinking about is in front of their eyes. In Wang et al 2020 it is made use of these findings and natural images were computationally discriminated based on gaze patterns during imagery and perception. We intend to use Virtual Reality as the environment of the experiment due to its potential to minimise the distractions that the participants in previous studies were subjected to. This may positively impact the accuracy of the gathered data. This study aims to investigate the use of reference frames, different classification and mapping pipelines as well as using more data generated by eye movement than in the previous study. The eye movements during recall are distorted compared to the ones made during encoding. Providing participants with reference frames during recall can minimise the above-mentioned distortions. The experimentation with different classification and mapping pipelines is meant to improve the performance of computational retrieval that so far reaches an accuracy of maximum of 72%.

What we did not mention:

We did not make completely clear what is the exact data we are intending to use further. The information gathered during the experiment is the gaze patterns you performed during the perception and imagery task together with their time stamps. It was particularly important for the integrity of the experiment not to disclose this information because people tend to exaggerate the eye movements, leading to unnatural behaviour when they know that their eye movements are monitored.

We remind you once again that you can withdraw from this study if you want to during the next two weeks.

Hypothesis:

It is hypothesised that images can be retrieved with high accuracy (above 70%) using eye movements during recall when the environment lacks distractions, a reference frame is provided, and the sequence of data is considered during classification.

Data Analysis:

Different processing approaches and classification techniques will be applied to the gaze patterns made during perception and visual imagery in order to computationally retrieve an image from a dataset.

Reading:

Computational discrimination between natural images based on gaze during mental imagery

https://www.nature.com/articles/s41598-020-69807-0#Sec15

If you have any follow up questions/want more details about the study, please send an email at amd77@bath.ac.uk.

12.6 12-Point Ethics Checklist

This form must be attached to the dissertation as an appendix.

Department of Computer Science

12-Point Ethics Checklist for UG and MSc Projects

Student Ana - Maria Dobre

Academic Year or Project Title Year 3

Supervisor Eamonn O'Neill

Does your project involve people for the collection of data other than you and your supervisor(s)? YES / NO

If the answer to the previous question is YES, you need to answer the following questions, otherwise you can ignore them.

This document describes the 12 issues that need to be considered carefully before students or staff involve other people ('participants' or 'volunteers') for the collection of information as part of their project or research. Replace the text beneath each question with a statement of how you address the issue in your project.

1. *Will you prepare a Participant Information Sheet for volunteers?* YES / NO

   Yes, a briefing script has been prepared such that the facilitator can brief participants accordingly.

2. *Will the participants be informed that they could withdraw at any time?* YES / NO

   Yes, participants will be able to withdraw from the study at any point.

3. *Will there be any intentional deception of the participants?* YES / NO

   Yes, participants will not be told that their eye movements are recorded because in previous similar studies, people tended to change their normal behaviour by exaggerating the movements leading to poor and unrealistic data sets. The participants will be told during the briefing that the pupil dimension is recorded and during de-briefing all information will be revealed clearly. At this point participants will be reminded that they can withdraw and their data be deleted if they do not agree with their data being used.

4. *Will participants be de-briefed?* YES / NO
   Yes, participants will be fully de-briefed after completing the study to protect the integrity of the research.

5. *Will participants voluntarily give informed consent?* YES / NO
   Yes, each participant will sign a consent form after being briefed.

6. *Will the participants be exposed to any risks greater than those encountered in their normal work life (e.g., through the use of non-standard equipment)?* YES / NO
   No, there is no serious risk posed to participants of this study. The only potential risk is experiencing motion-sickness as a result of using a virtual reality headset, however, this will be mitigated by planning frequent breaks, and allowing participants to withdraw at any point.

7. *Will you be offering any incentive to the participants?* YES / NO
   No, participants will not receive any payment.

8. *Will you be in a position of authority or influence over any of your participants?* YES / NO
   No, I am an undergraduate student at the University of Bath.

9. *Will any of your participants be under the age of 16?* YES / NO
   No, all participants will be over 16.

10. *Will any of your participants have an impairment that will limit Their understanding or communication?* YES / NO
    No, participants with limiting impairments will not be included in the study.

11. *Will the participants be informed of your contact details?* YES / NO
    Yes, all participants will be provided with contact details.

12. *Will you have a data management plan for all recorded data?* YES / NO
    Yes, all data will be stored securely on an encrypted hard drive.

12.7 EIRA1 form

FORM A: ETHICS REVIEW CHECKLIST FOR RESEARCH WITH HUMAN PARTICIPANTS

This checklist should be completed for every research project involving human participants in order to identify whether a full application for ethics approval needs to be submitted to one of the University Ethics Sub-Committees. The principal investigator or, where the principal investigator is a student, the supervisor, is responsible for exercising appropriate professional judgement in this review.

Section 1: Project details			
Project title:	Prediction of perceived complex visual stimuli based on eye movements during perception and visual imagery in a VR environment		
Brief synopsis of study: (no more than 250 words)	This study aims to improve on a previous research conducted by Wang et al in 2020 where a natural images were computationally discriminated based on gaze patterns during imagery and perception. We intend to use Virtual Reality as environment due to its potential to minimise the distractions that the participants in previous studies were subjected to. This may positively affect the accuracy of the gathered data. This study aims to investigate the use of reference frames, different classification and mapping pipelines as well as		
Planned start date:	10th of October 2020	Planned end date:	30th of April 2021
Funder:	no		
Do you have prior approval?	no		
If so, please state where from			

Section 2: Applicant details			
Applicant name and username:	Ana-Maria Dobre amd77		
Department:	Computer Science		
Email:	amd77@bath.ac.uk		
Undergraduate ☒	Masters ☐	Research Postgraduate ☐	Staff ☐

(NB: If you have prior ethical approval from the NHS, SCREC or a UK academic institution, please sign this checklist and attach evidence of that approval. There is no need to complete any more questions.)

3(A) Research that may need full review by a University of Bath Ethics Sub-Committee	**YES**	**NO**
Do you and/or your supervisor intend to submit your results for publication to the wider research community (other than in your dissertation) where evidence of ethical approval is required?	☐	☒
Does the project involve the collection of material that could be considered of a personal, biographical, medical, psychological, social or physiological nature?	☐	☒
Does the research involve vulnerable groups: e.g. children; those with cognitive impairment; or those in unequal relationships (e.g. your own students)?	☐	☒
Will the study require the cooperation of a gatekeeper for initial access to the groups or individuals to be recruited (e.g. headmaster at a School; group leader of a self-help group)?	☐	☒
Will it be necessary for participants to take part in the study without their knowledge and consent at the time? (e.g. covert observation of people in non-public places?)	☐	☒
Will the study involve discussion of sensitive topics (e.g. sexual activity; drug use; criminal activity)?	☐	☒
Is pain or more than mild discomfort likely to result from the study?	☐	☒
Could the study induce psychological stress or anxiety or cause harm or negative consequences beyond the risks encountered in normal life?	☐	☒
Will the study involve prolonged or repetitive testing?	☐	☒

FORM A: ETHICS REVIEW CHECKLIST FOR RESEARCH WITH HUMAN PARTICIPANTS – FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Will the research involve organisational administrative or secure data that requires permission from the appropriate organisation/authorities before use (data that is not in the public domain)?	☐	☒
Does the research involve participants carrying out any of the research activities themselves (i.e. acting as researchers as opposed to just being participants)?	☐	☒
Is there a possibility that the safety of the researcher may be in question (e.g. international research; locally employed research assistants)?	☐	☒
Will personal data be transferred to the UK from outside the EEA?	☐	☒
Will the outcome of the research allow respondents to be identified either directly or indirectly (e.g. through aggregating separate data sources gathered from the internet)?	☐	☒
Will research involve the sharing of data or confidential information beyond the initial consent given?	☐	☒
Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?	☐	☒
Will the proposed findings be controversial or are there any conflicts of interest?	☐	☒
Will the study involve the publication, sharing or potentially insecure electronic storage and/or transfer of data that might allow identification of individuals, either directly or indirectly? (e.g. publication of verbatim quotations from an online forum; sharing of audio/visual recordings; insecure transfer of personal data such as addresses, telephone numbers etc.; collecting identifiable personal data on unprotected** internet sites.) [**Please note that Qualtrics provides adequate data security and comply with the requirements of the EU-US Privacy Shield.]	☐	☒

3(B) Security Sensitive Material	**YES**	**NO**
Does your research involve access to or use of material covered by the Terrorism Act? (The Terrorism Act (2006) outlaws the dissemination of records, statements and other documents that can be interpreted as promoting and endorsing terrorist acts. By answering 'yes' you are registering your legitimate use of this material with the Research Ethics Advisory Group. In the event of a police investigation, this registration will help you to demonstrate that your use of this material is legitimate and lawful).	☐	☒

3(C) Prevent Agenda	**YES**	**NO**
Does the research have the potential to radicalise people who are vulnerable to supporting terrorism or becoming terrorists themselves?	☐	☐

If the answer to <u>all</u> questions in Sections 3(A) and/or 3(B) and/or 3(C) is 'no', please complete a an EIRA1 submission for review within the Department of Computer Science, and attach this form to your EIRA1 submission.

If the answer to <u>any</u> questions in Sections 3(A) and/or 3(B) and/or 3(C) is 'yes', please complete the full application form for an appropriate University Ethics Sub-Committee and send it to the sub-committee secretary, together with required supporting documentation.

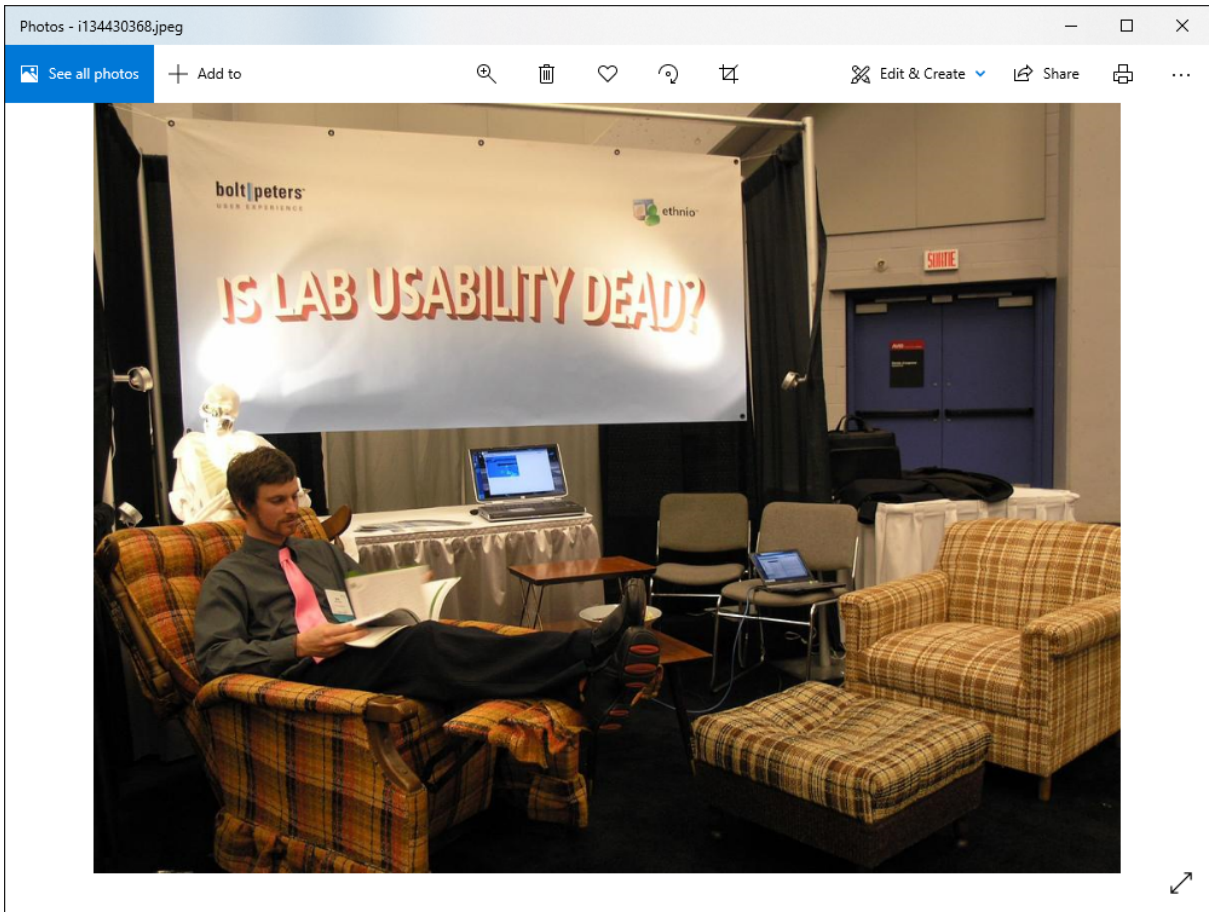
Section 4: Declaration and signatures			
Please note that it is your responsibility to follow, and to ensure that, all researchers involved with your project follow accepted ethical practice and appropriate professional ethical guidelines in the conduct of your study. You must take all reasonable steps to protect the dignity, rights, safety and well-being of participants. This includes providing participants with appropriate information sheets, ensuring informed consent and ensuring confidentiality in the storage and use of data.			
Applicant signature	Ana Dobre	**Date**	30.11.2020

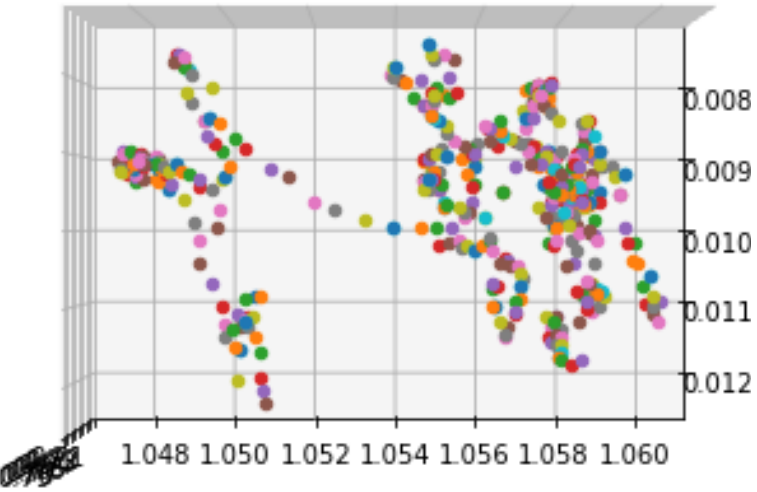
Supervisor name	Eamonn O'Neill	**Date**	

FORM A: ETHICS REVIEW CHECKLIST FOR RESEARCH WITH HUMAN PARTICIPANTS – FACULTY OF HUMANITIES AND SOCIAL SCIENCES

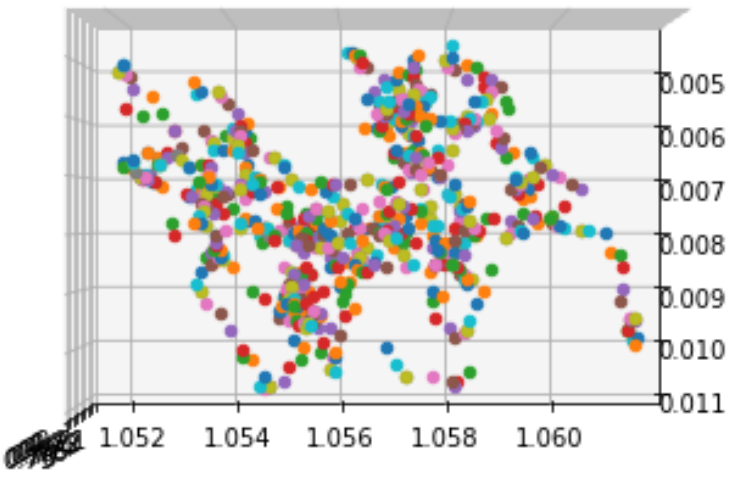
Supervisor signature			

12.8 Preliminary Tests

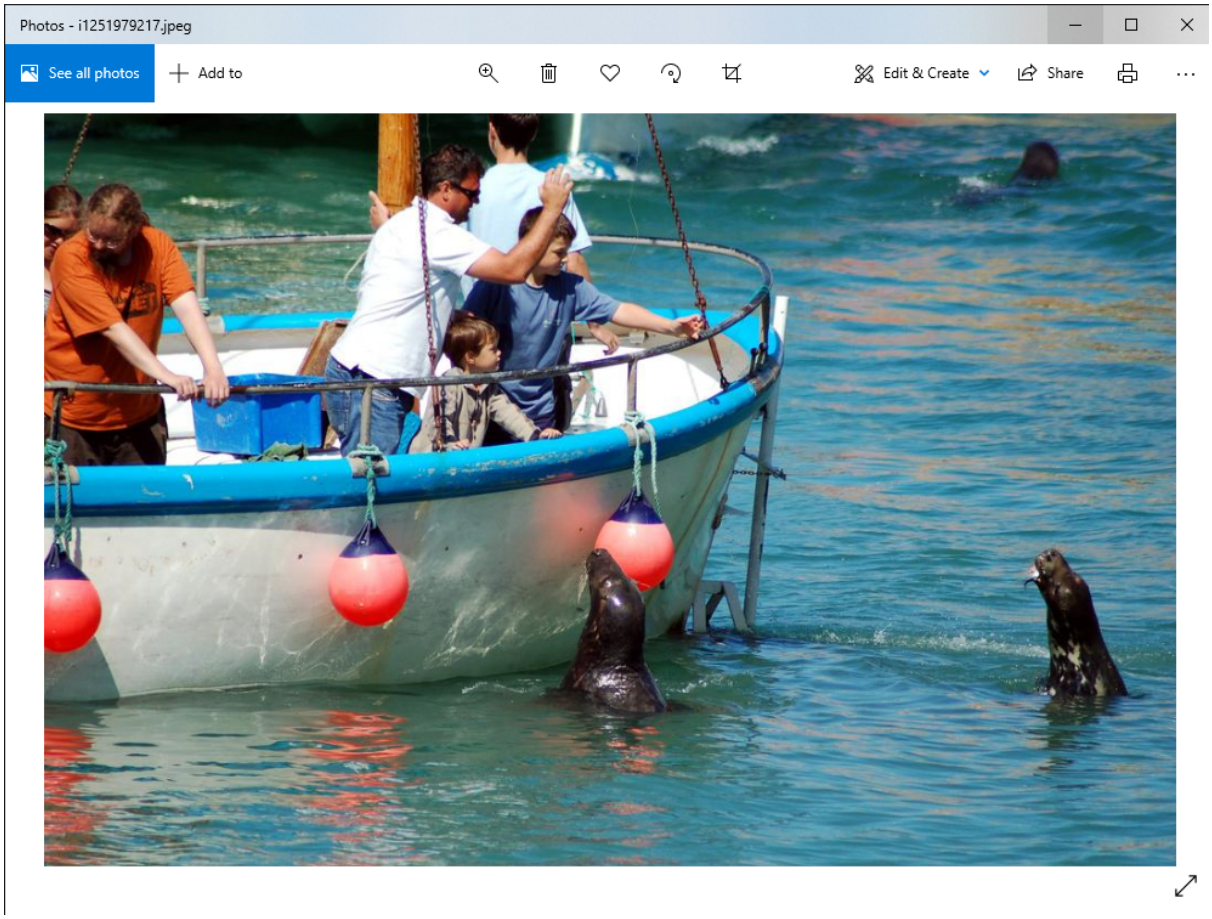


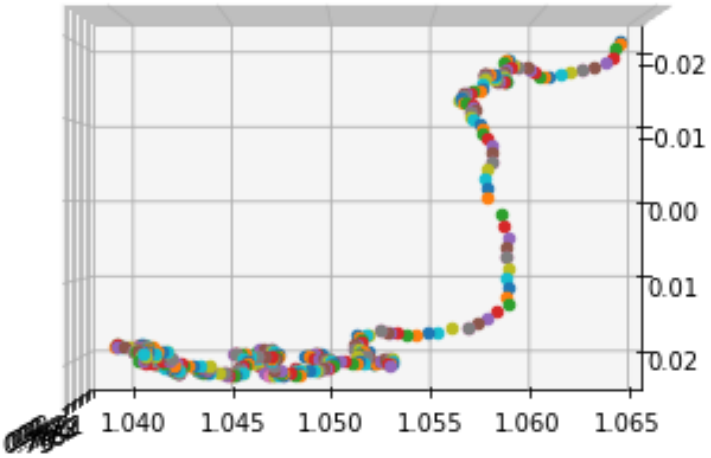


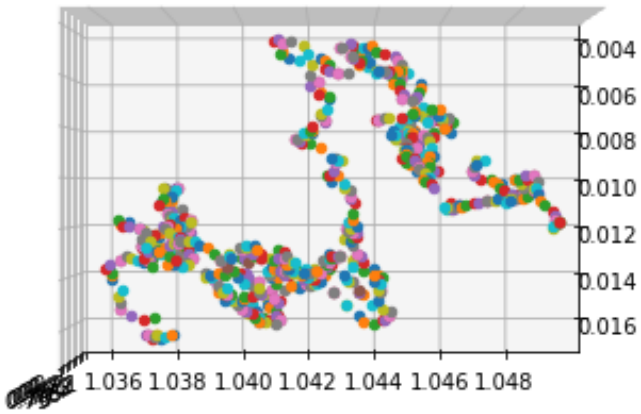
Encoding



Recall



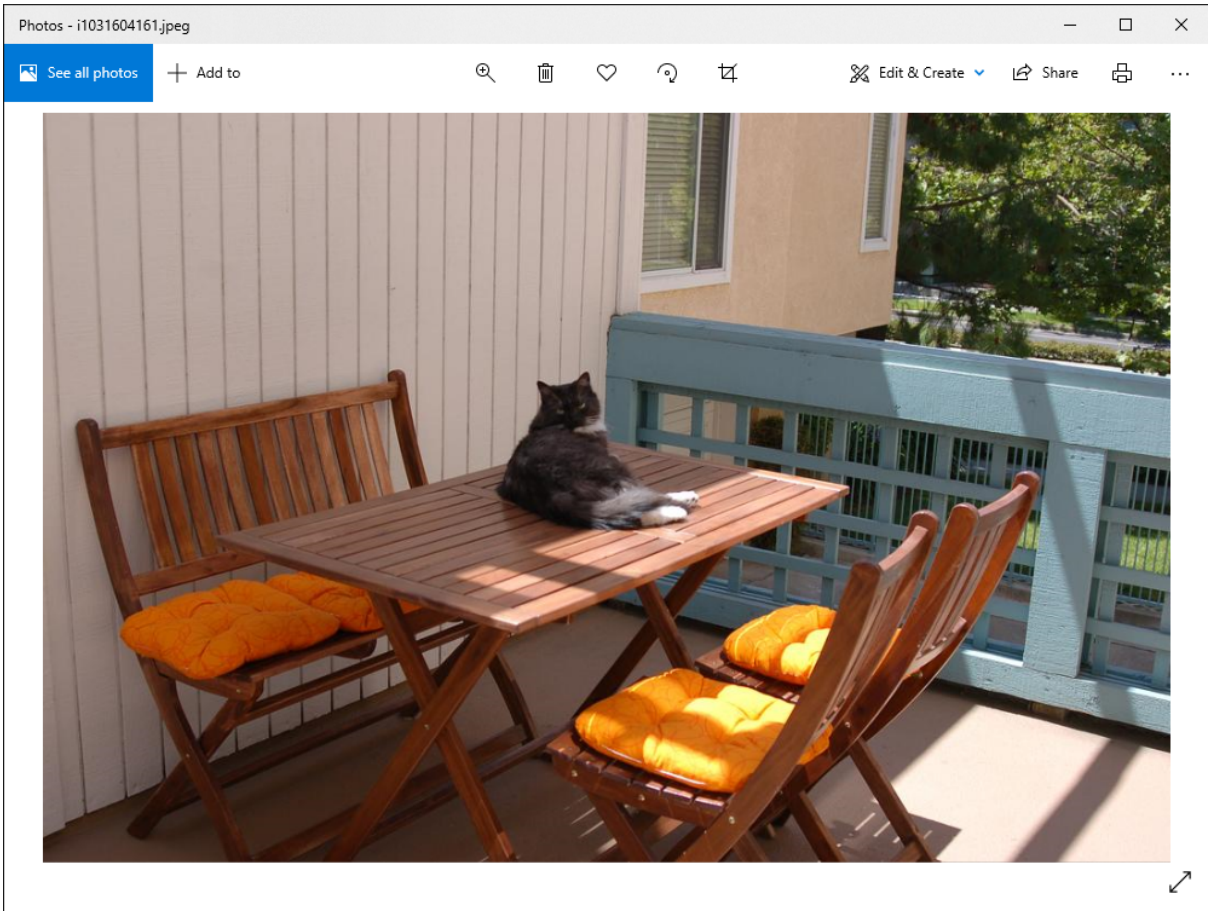


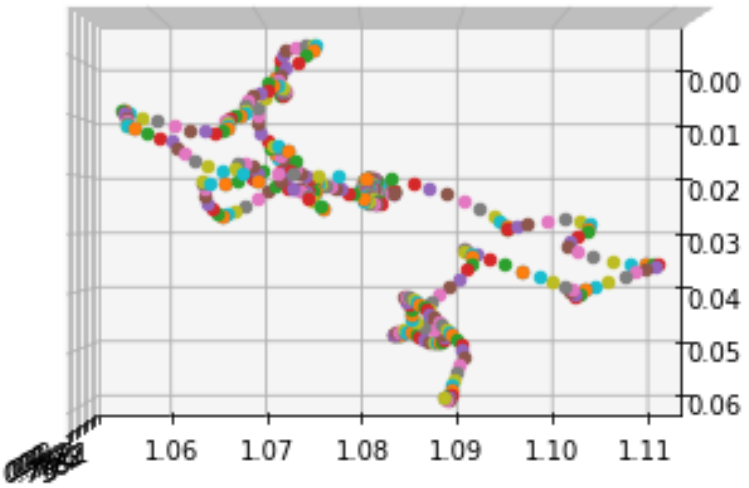


Encoding

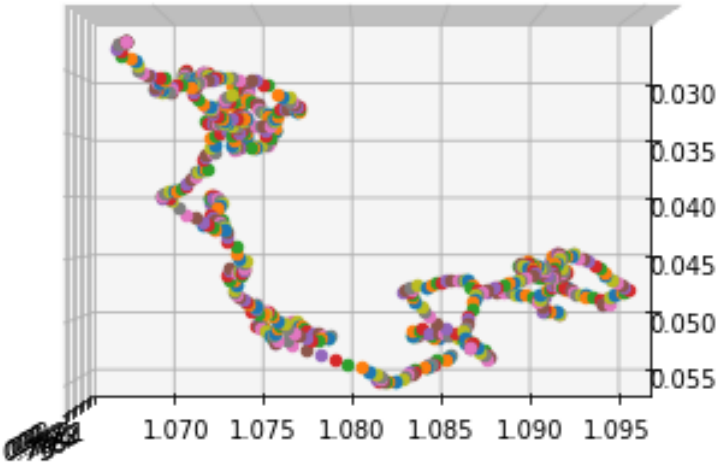
Recall

if I would disregard the data between [1.055, 1.065] and [-0.02, 0.015] the encoding data would look better. not sure what the participant did there.

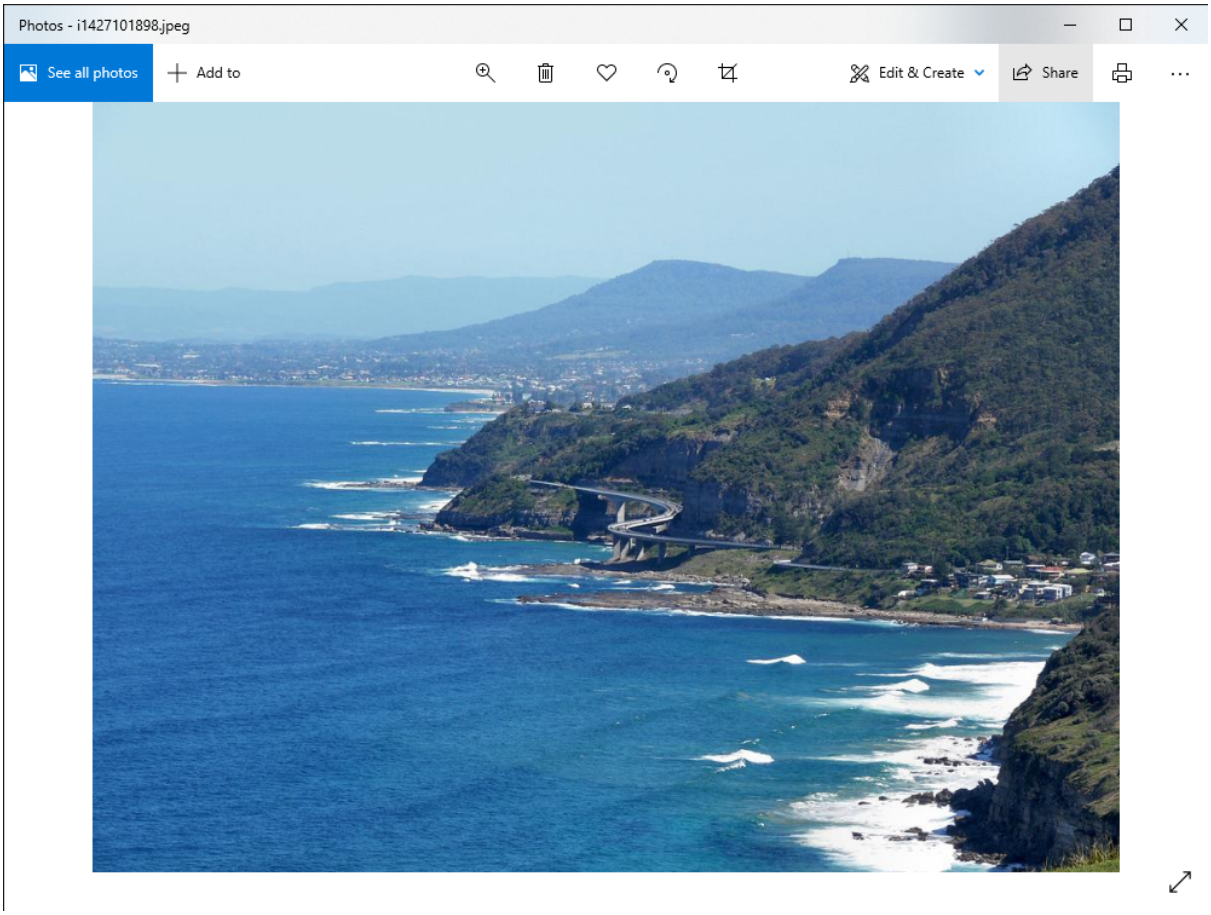


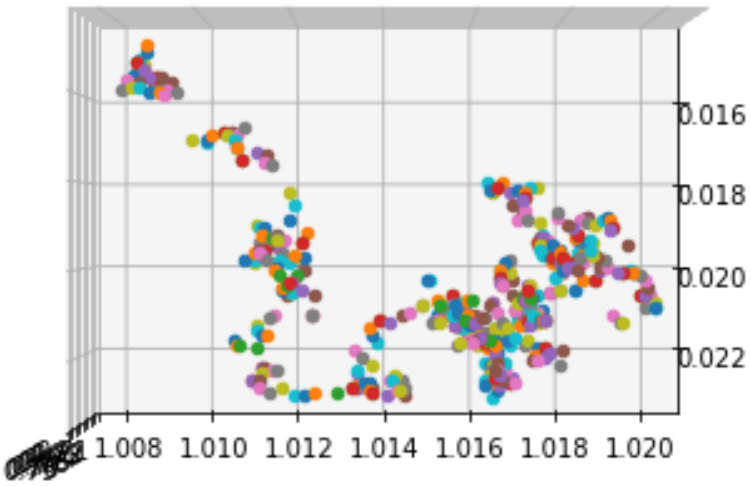


Encoding

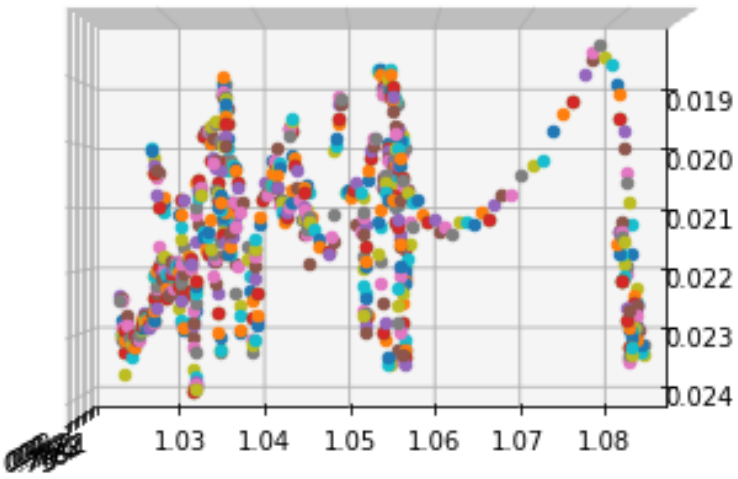


Recall

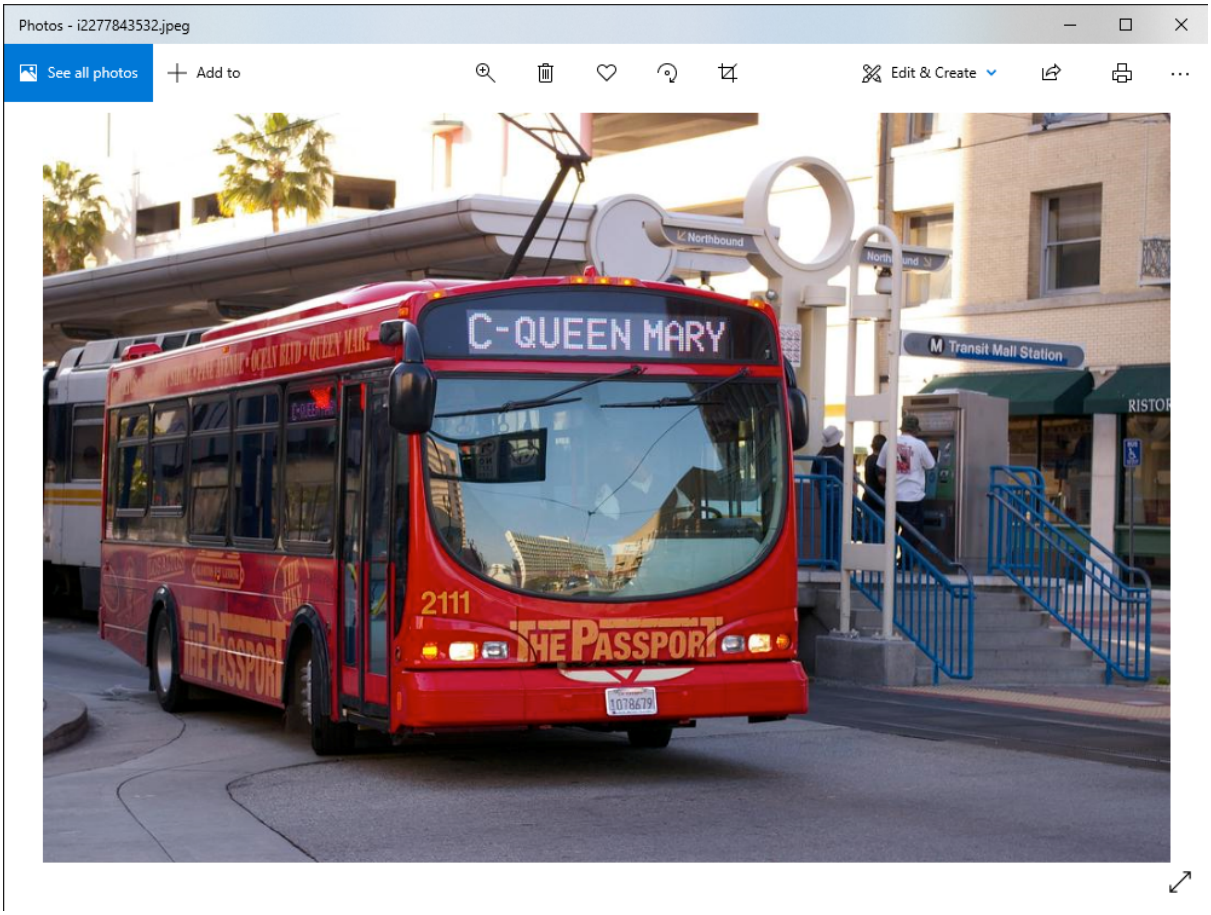


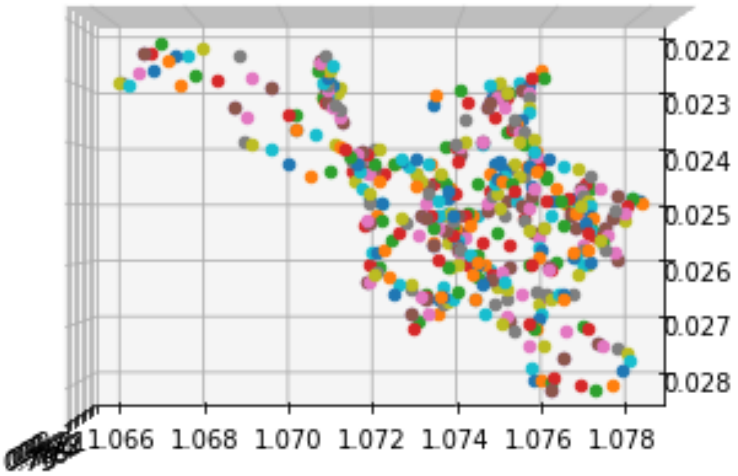


Encoding

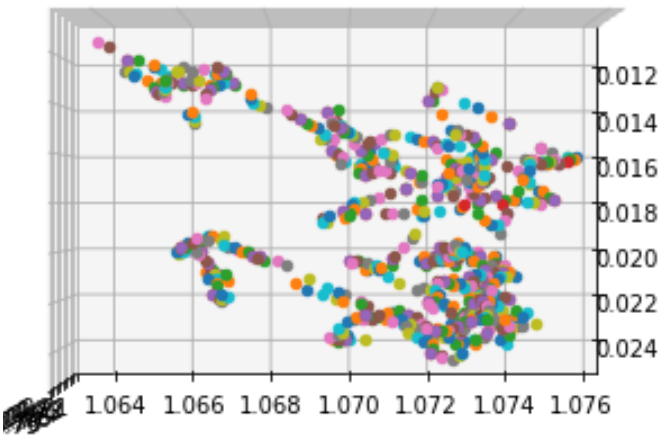


Recall

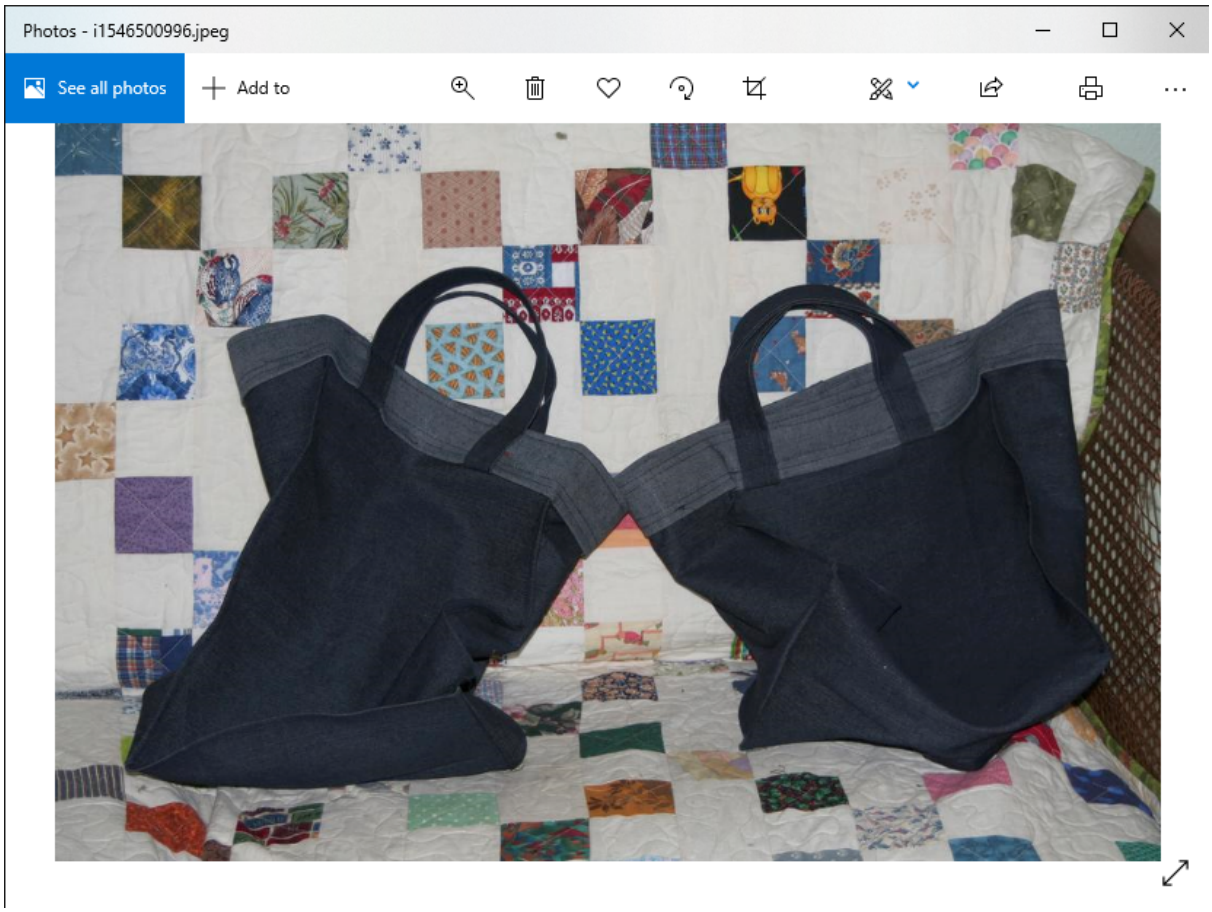


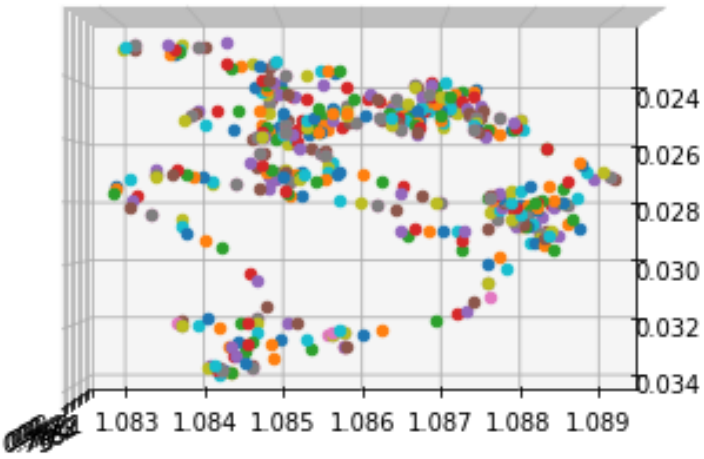


Encoding

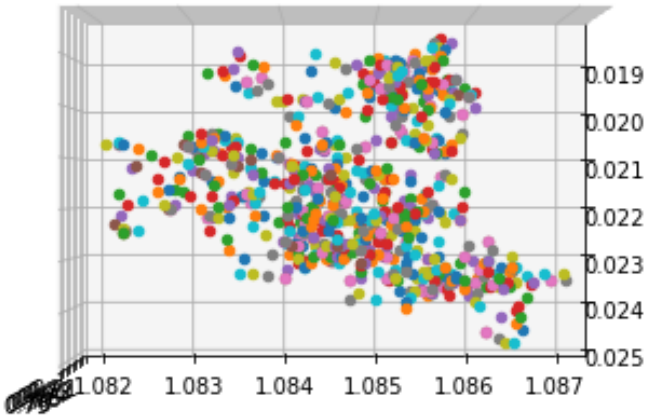


Recall

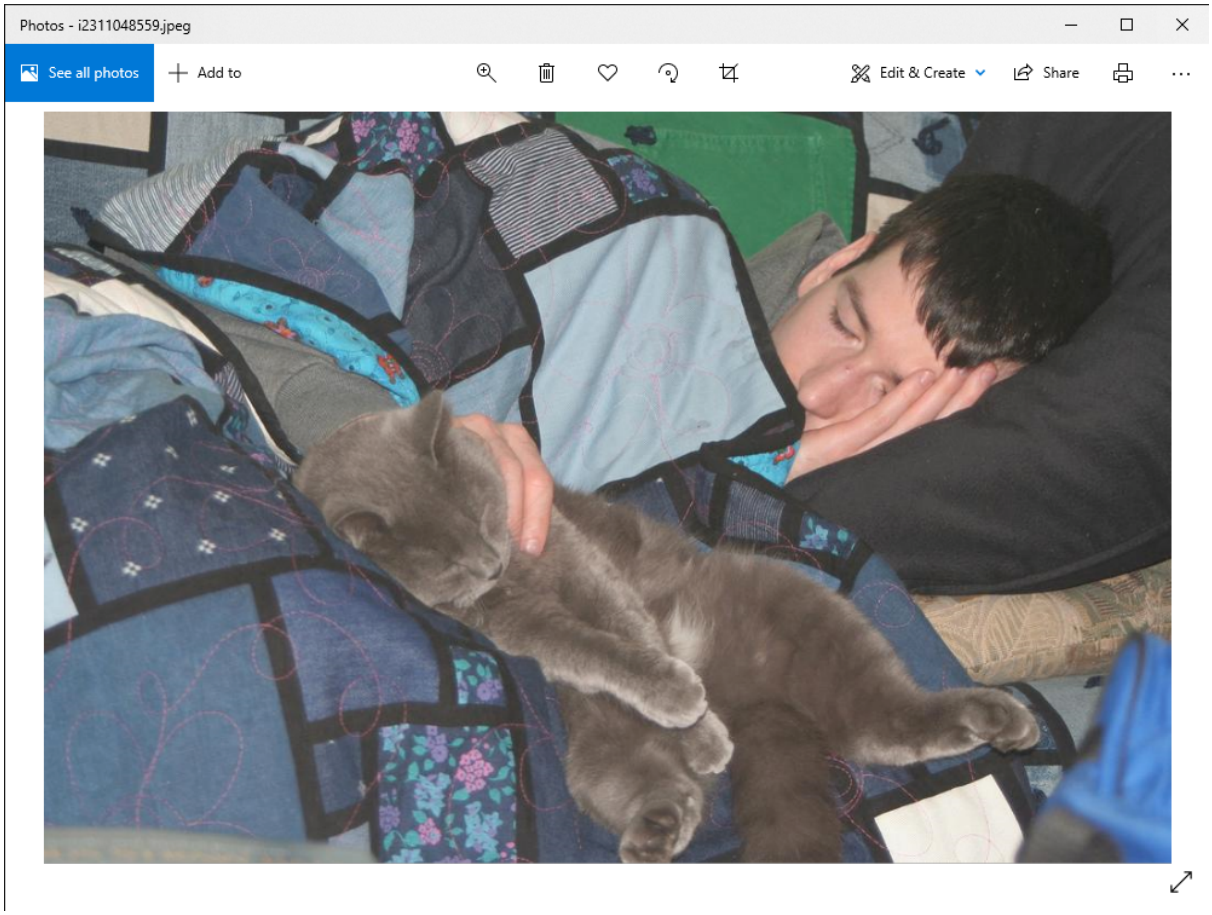


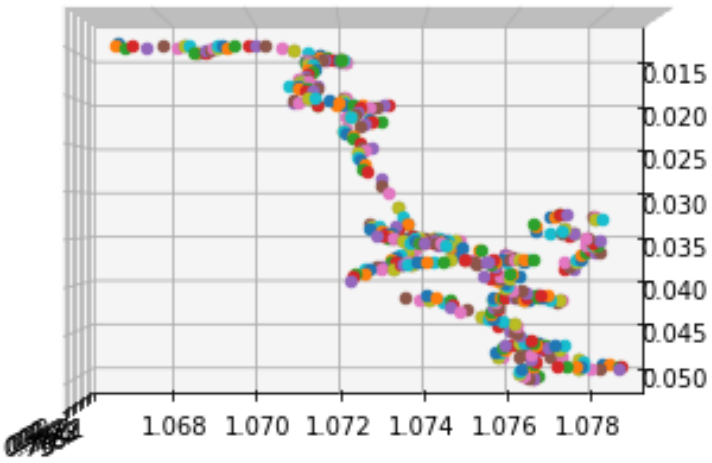


Encoding

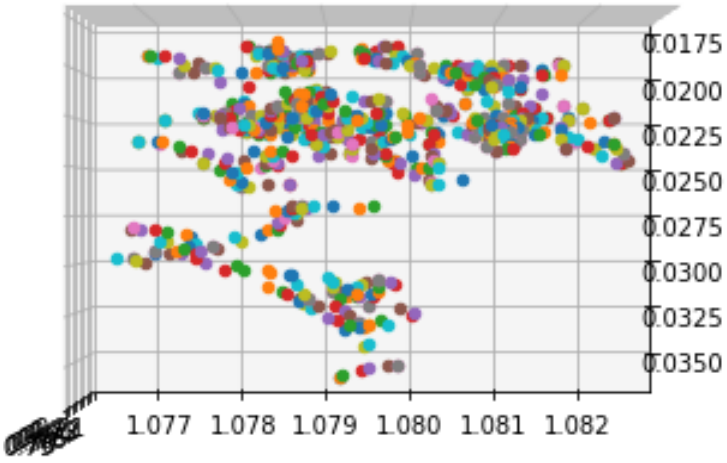


Recall

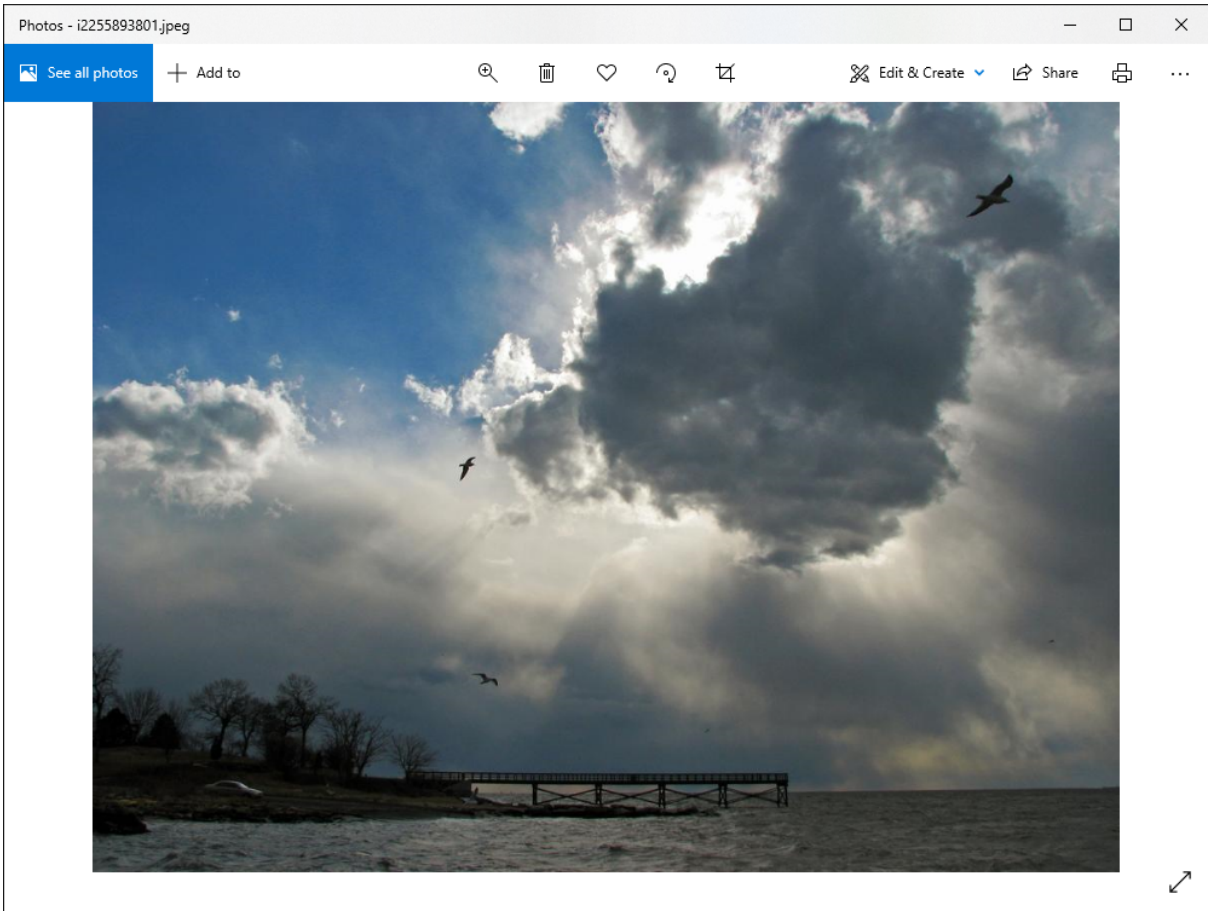


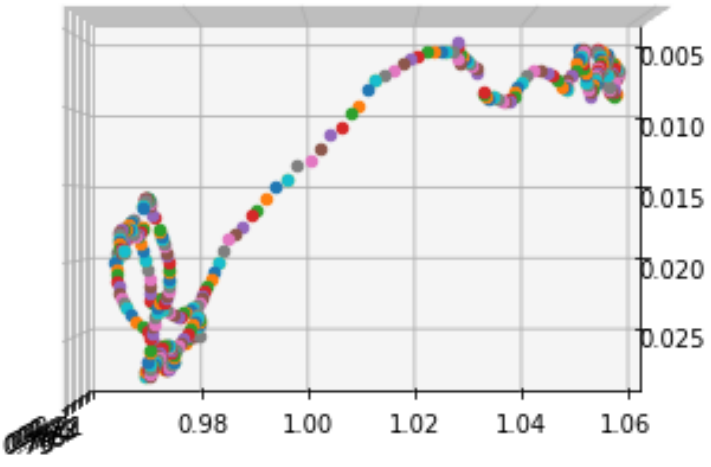


Encoding

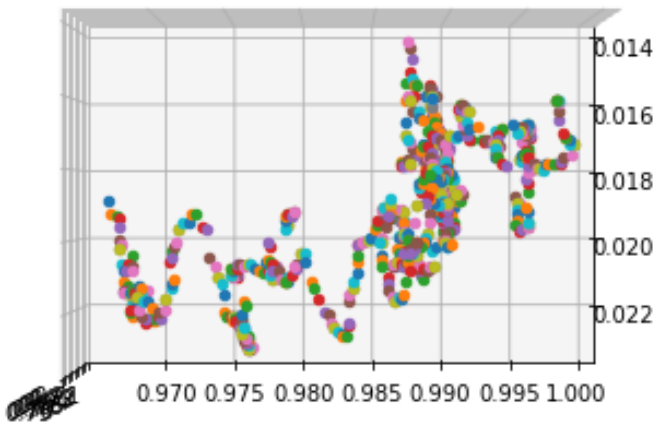


Recall

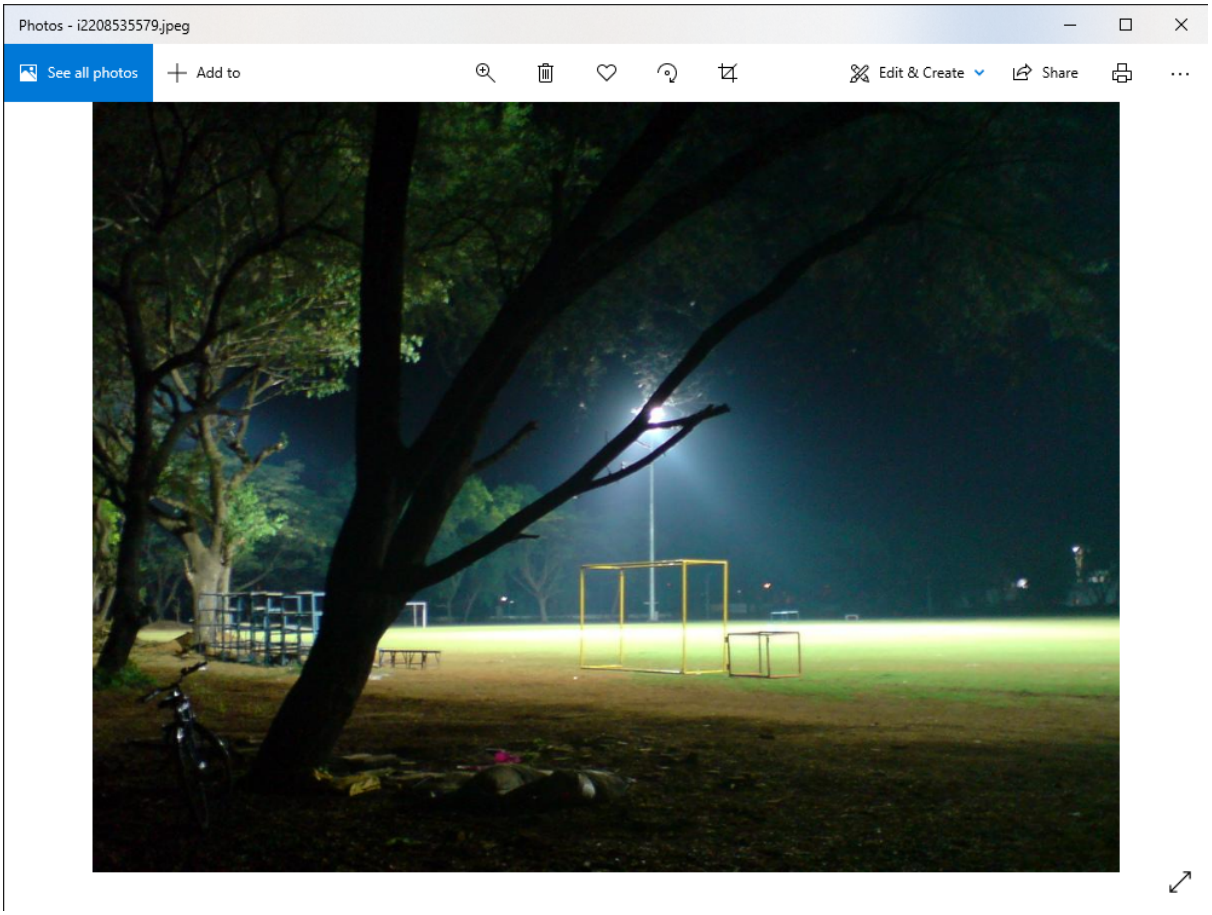


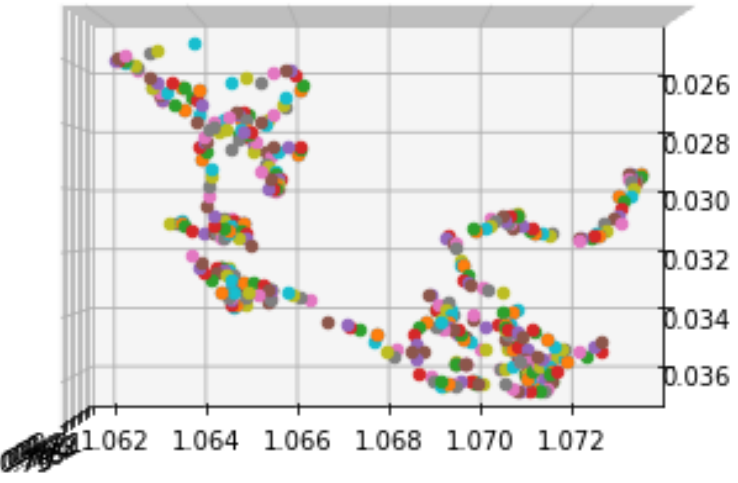


Encoding

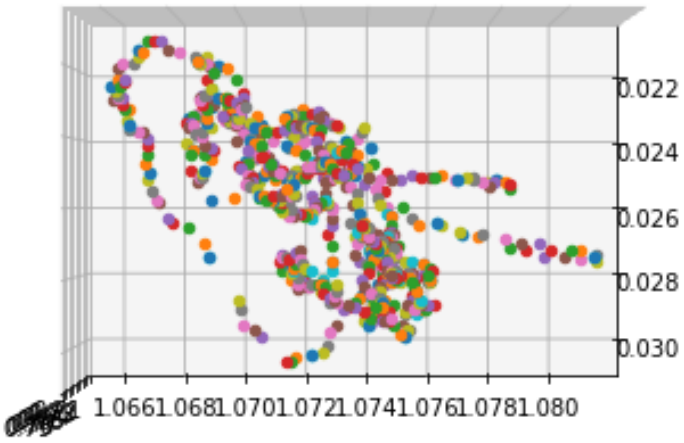


Recall





Encoding



Recall

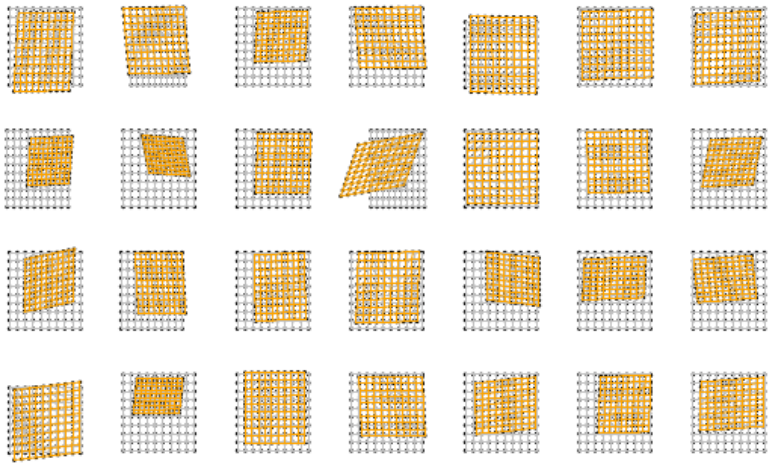


Figure 1.3.2. Example of distorted gaze patterns during recall (yellow) compared to gaze patterns during encoding (gray)(Wang et al., 2020)

Bibliography

Barnes, G. R. (2008), 'Cognitive processes involved in smooth pursuit eye movements', *Brain and cognition* **68**(3), 309–326.

Bay, H., Tuytelaars, T. and Van Gool, L. (2006), Surf: Speeded up robust features, *in* 'Proceedings of the 9th European Conference on Computer Vision - Volume Part I', ECCV'06, Springer-Verlag, Berlin, Heidelberg, p. 404–417.
URL: *https://doi.org/10.1007/11744023*$_3$2

Behrmann, M. (2000), 'The mind's eye mapped onto the brain's matter', *Current Directions in Psychological Science* **9**(2), 50–54.

Blazhenkova, O. and Kozhevnikov, M. (2009), 'The new object-spatial-verbal cognitive style model: Theory and measurement', *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* **23**(5), 638–663.

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.

Branson, S., Van Horn, G., Wah, C., Perona, P. and Belongie, S. (2014), 'The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization', *International Journal of Computer Vision* **108**(1-2), 3–29.

Chadwick, M. J., Hassabis, D., Weiskopf, N. and Maguire, E. A. (2010), 'Decoding individual episodic memory traces in the human hippocampus', *Current Biology* **20**(6), 544–547.

Chaudhary, U., Birbaumer, N. and Ramos-Murguialday, A. (2016), 'Brain–computer interfaces for communication and rehabilitation', *Nature Reviews Neurology* **12**(9), 513.

Coddington, J., Xu, J., Sridharan, S., Rege, M. and Bailey, R. (2012), Gaze-based image retrieval system using dual eye-trackers, *in* '2012 IEEE International Conference on Emerging Signal Processing Applications', IEEE, pp. 37–40.

Cowen, A. S., Chun, M. M. and Kuhl, B. A. (2014), 'Neural portraits of perception: reconstructing face images from evoked brain activity', *Neuroimage* **94**, 12–22.

Daly, J. J. and Wolpaw, J. R. (2008), 'Brain–computer interfaces in neurological rehabilitation', *The Lancet Neurology* **7**(11), 1032–1043.

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.

Eger, N., Ball, L. J., Stevens, R. and Dodd, J. (2007), Cueing retrospective verbal reports in usability testing through eye-movement replay, *in* 'Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21', pp. 1–9.

Faro, A., Giordano, D., Pino, C. and Spampinato, C. (2010), Visual attention for implicit relevance feedback in a content based image retrieval, *in* 'Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications', pp. 73–76.

Gentaz, E. and Hatwell, Y. (2004), 'Geometrical haptic illusions: The role of exploration in the müller-lyer, vertical-horizontal, and delboeuf illusions', *Psychonomic Bulletin & Review* **11**(1), 31–40.

Goldenshluger, A., Zeevi, A. et al. (2004), 'The hough transform estimator', *The Annals of Statistics* **32**(5), 1908–1932.

Hamamé, C. M., Vidal, J. R., Ossandón, T., Jerbi, K., Dalal, S. S., Minotti, L., Bertrand, O., Kahane, P. and Lachaux, J.-P. (2012), 'Reading the mind's eye: online detection of visuo-spatial working memory and visual imagery in the inferior temporal lobe', *Neuroimage* **59**(1), 872–879.

Hayhoe, M. M. (2017), 'Vision and action', *Annual review of vision science* **3**, 389–413.

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.

Hornsey, R. L., Hibbard, P. B. and Scarfe, P. (2020), 'Size and shape constancy in consumer virtual reality', *Behavior research methods* **52**, 1587–1598.

Hosu, V., Lin, H., Sziranyi, T. and Saupe, D. (2020), 'Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment', *IEEE Transactions on Image Processing* **29**, 4041–4056.

Jacobson, E. (1932), 'Electrophysiology of mental activities', *The American Journal of Psychology* **44**(4), 677–694.

Jerbi, K., Freyermuth, S., Minotti, L., Kahane, P., Berthoz, A. and Lachaux, J.-P. (2009), 'Watching brain tv and playing brain ball: Exploring novel bci strategies using real-time analysis of human intracranial data', *International review of neurobiology* **86**, 159–168.

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. and Homqvist, K. (2011), The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 33.

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.

Johansson, R. and Johansson, M. (2014), 'Look here, eye movements play a functional role in memory retrieval', *Psychological Science* **25**(1), 236–242.

Judd, T., Ehinger, K., Durand, F. and Torralba, A. (2009), Learning to predict where humans look, *in* '2009 IEEE 12th International Conference on Computer Vision', pp. 2106–2113.

Karami, E., Shehata, M. and Smith, A. (2017), 'Image identification using sift algorithm: Performance analysis against different image deformations', *arXiv preprint arXiv:1710.02728* .

Kim, S.-P., Simeral, J. D., Hochberg, L. R., Donoghue, J. P. and Black, M. J. (2008), 'Neural control of computer cursor velocity by decoding motor cortical spiking activity in humans with tetraplegia', *Journal of neural engineering* **5**(4), 455.

Klatzky, R. (1998), 'Spatial cognition-an interdisciplinary approach to representation and processing of spatial knowledge'.

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., Bloem, I. M., D'Ascenzo, S. and Tommasi, L. (2014), 'Scrutinizing visual images: The role of gaze in mental imagery and memory', *Cognition* **131**(2), 263–283.

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* .

Liversedge, S. P. and Findlay, J. M. (2000), 'Saccadic eye movements and cognition', *Trends in cognitive sciences* **4**(1), 6–14.

Marks, D. F. (1973), 'Visual imagery differences and eye movements in the recall of pictures', *Perception & psychophysics* **14**(3), 407–412.

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.

Mirza, S. N. H., Proulx, M. J. and Izquierdo, E. (2012), 'Reading users' minds from their eyes: A method for implicit image annotation.', *IEEE Trans. Multimedia* **14**(3-2), 805–815.

Moore, C. S. (1903), 'Control of the memory image.', *The Psychological Review: Monograph Supplements* .

Mulder, T., Zijlstra, S., Zijlstra, W. and Hochstenbach, J. (2004), 'The role of motor imagery in learning a totally novel movement', *Experimental brain research* **154**(2), 211–217.

Perky, C. W. (1910), 'An experimental study of imagination', *The American Journal of Psychology* **21**(3), 422–452.

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.

Rayner, K. (1998), 'Eye movements in reading and information processing: 20 years of research.', *Psychological bulletin* **124**(3), 372.

Rayner, K., Smith, T. J., Malcolm, G. L. and Henderson, J. M. (2009), 'Eye movements and visual encoding during scene perception', *Psychological science* **20**(1), 6–10.

Ribelles, J., Gutierrez, D. and Efros, A. (2017), 'Buildup: interactive creation of urban scenes from large photo collections', *Multimedia Tools and Applications* **76**(10), 12757–12774.

Richardson, A. (2013), *Mental imagery*, Springer.

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.

Rosten, E. and Drummond, T. (2006), Machine learning for high-speed corner detection, *in* 'European conference on computer vision', Springer, pp. 430–443.

Schalk, G., Kubanek, J., Miller, K., Anderson, N., Leuthardt, E., Ojemann, J., Limbrick, D., Moran, D., Gerhardt, L. and Wolpaw, J. (2007), 'Decoding two-dimensional movement trajectories using electrocorticographic signals in humans', *Journal of neural engineering* **4**(3), 264.

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.

Schütz, A. C., Braun, D. I. and Gegenfurtner, K. R. (2011), 'Eye movements and perception: A selective review', *Journal of vision* **11**(5), 9–9.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T. and Kamitani, Y. (2019*a*), 'End-to-end deep image reconstruction from human brain activity', *Frontiers in Computational Neuroscience* **13**, 21.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T. and Kamitani, Y. (2019*b*), '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*

Steichen, B., Wu, M. M., Toker, D., Conati, C. and Carenini, G. (2014), Te, te, hi, hi: Eye gaze sequence analysis for informing user-adaptive information visualizations, *in* 'International Conference on User Modeling, Adaptation, and Personalization', Springer, pp. 183–194.

Tan, D. and Nijholt, A. (2010), Brain-computer interfaces and human-computer interaction, *in* 'Brain-Computer Interfaces', Springer, pp. 3–19.

Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. and van Lier, R. (2019), 'Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience', *Scientific reports* **9**(1), 1–8.

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**.

Tirupattur, P., Rawat, Y. S., Spampinato, C. and Shah, M. (2018), Thoughtviz: Visualizing human thoughts using generative adversarial network, *in* 'Proceedings of the 26th ACM international conference on Multimedia', pp. 950–958.

Tripathy, S., Kannala, J. and Rahtu, E. (2018), 'Learning image-to-image translation using paired and unpaired training samples'.

Tsai, F. C. (1994), 'Geometric hashing with line features', *Pattern Recognition* **27**(3), 377–389.

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* .

van den Boom, M. A., Vansteensel, M. J., Koppeschaar, M. I., Raemaekers, M. A. and Ramsey, N. F. (2019), 'Towards an intuitive communication-bci:

decoding visually imagined characters from the early visual cortex using high-field fmri', *Biomedical Physics & Engineering Express* **5**(5), 055001.

Van Gerven, M., Bahramisharif, A., Heskes, T. and Jensen, O. (2009), 'Selecting features for bci control based on a covert spatial attention paradigm', *Neural Networks* **22**(9), 1271–1277.

Van Gerven, M. and Jensen, O. (2009), 'Attention modulations of posterior alpha as a control signal for two-dimensional brain–computer interfaces', *Journal of neuroscience methods* **179**(1), 78–84.

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, R. (2007), 'Spatial processing in navigation, imagery, and perception'.

Wang, R. F. (2012), 'Theories of spatial representations and reference frames: What can configuration errors tell us?', *Psychonomic bulletin & review* **19**(4), 575–587.

Wang, X., Bylinskii, Z., Castelhano, M., Hillis, J. and Duchowski, A. T. (2020*a*), Emics'20: Eye movements as an interface to cognitive state, *in* 'Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems', pp. 1–4.

Wang, X., Bylinskii, Z., Castelhano, M., Hillis, J. and Duchowski, A. T. (2020*b*), Emics'20: Eye movements as an interface to cognitive state, *in* 'Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems', CHI EA '20, Association for Computing Machinery, New York, NY, USA, p. 1–4.
URL: *https://doi.org/10.1145/3334480.3381062*

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.

Yarbus, A. (1967), 'Eye movements during perception of complex objects [online, dostep: 2016-02-01]. w: Al yarbus. eye movements and vision (s. 171–211)'.

Zhou, Y., Wang, J. and Chi, Z. (2018), Content-based image retrieval based on eye-tracking, *in* 'Proceedings of the Workshop on Communication by Gaze Interaction', pp. 1–7.

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`