Literature and Technology Survey
Prediction of imagined and perceived complex visual stimuli in a VR environment based on eye movements

Ana Dobre

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

Contents

Chapter 1

Introduction

1.1 Problem Description

This project aims to improve on performance of a recent Nature publication (Wang et al., 2020), which attempted to computationally discriminate a visual image amongst 100 images based on gaze patterns from observers perceiving those images, and apply these classifiers to gaze data from recall and visual imagery of the same stimuli. Whilst Wang et al (2020) achieved impressive results for discriminating perceived images using K-Nearest Neighbour and CNN classifiers, performance for visual imagery was poor. Low performance is likely driven by several factors such as: scaled, shifted and translated gaze patterns during imagery, compared to perception. Gaze data was reduced into histograms, reducing discriminability between pictures with similar histograms. Participants had their eyes open which resulted in perceptual information interference during the imagery task. The method differences for this project are the usage of a virtual reality environment and a different computational design to improve performance.

The essential question in Wang et al's 2020b study was "how well can images be computationally discriminated from other images based on only gaze data". Even though previous studies demonstrated that encoding and recall eye movements are similar, it was not clear how well the later ones performed as the input data for computational discrimination of an image. Wang et al., 2020b was the first paper to attempt computational retrieval using gaze patterns from visual imagery. There is a vast body of literature on fMRI and EEG based classification and reconstruction of perception (and to a 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). Wang et al., 2020b made the process of image retrieval using eye movement patterns more transparent compared to previous implementations in the field of CBIR.

1.2 Motivation

The problem of computationally discriminating a visual stimulus from a dataset based on encoding and recall eye data is worth studying for multiple reasons. Firstly, 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 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. 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).

Secondly, 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). Wang et al., 2020a explains in a paper presented during CHI 2020 that eye movement can be used to determine the intentions of users, difficulty of tasks etc., and shape the user's interaction with systems (guide attention, affect processing, and provide aids for learning and memory). Recalling is a part of visual imagery and it represents a cognitive state that can be tracked and analysed through involuntary eye movements (Wang et al., 2020b). 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.

Chapter 2

Literature and Technology Survey

2.1 Identify different methods for retrieving a mental image

Brain-computer interfaces

Image retrieval can be tackled as a classification problem where some cues (such as brain signals or eye movements) represent the input for a computational system that classifies them as a representation of a visual stimuli 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. 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 an imagined visual stimuli has had very little success so far.

Limitations of BCI

Even though fMRI and EEG based BCIs show impressive results when classifying perception or visual imagery, they impose numerous motor limitations. Even though BCIs usually aim to help people with locked-in syndrome (LIS), the lack of mobility imposed on participants by the fMRI and EEG machines makes the wide implementation of technologies that use perception or visual imagery imposible. 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.

Current gaze tracking technology and the strong evidence that during imagery people perform eye movements that resemble the ones made during encoding could address this BCI limitations and offer a more innately efficient solution.

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. 2020b 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 subjects can result in similar histograms, making the process of retrieval difficult for large databases (especially when discrete histograms are used. This proved to be an issue even for a database of 100 natural images. 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 2020b suggests, 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 encode 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. 2020b. 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 fovera where cones are concentrated and the highest resolution is achieved and therefore, fine details can be observed. Eye movements are usually directed exactly to the stimuli that catches one's attention unless covert attention is at play (Holmqvist et al., 2011).

Mulder et al. 2004 and Jacobson 1932 examined eye movements during perception and imagery and concluded that it is only muscle that shows a similar amount of activation. 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 retrieval to new images would be extremely difficult without intrinsic similarity between eye movements during encoding and recall. Wang et al. 2020b 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.

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 behaviors 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. Naturally, they correspond to the desire to maintain one's gaze on an object of interest. 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. 2020b 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% recall fixations were located inside of the stimuli domain, while 99% of encoding fixations were within the stimuli boundaries in Wang et al 2020b. 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. 2020b 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. Adding time series to the histograms could represent one way of computing less similar input data for different visual stimuli 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. 2020b 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". As expected, 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. The Nature paper only gave 5 seconds. Wang et al., 2020 suggests that the 5s only give enough time for a participant to recall the very basic layout of the stimuli. Exploring different representations for this type of data 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. In order to implement such changes, the recall time might need to be extended. Eye movements sequence can also reflect the image memorability. Gaze patterns partially highlight the visual features a viewer is driven to. Linsley et al., 2019 and Yang et al., 2016 could guide the integration of image content in future work. Image memorability (Yang et al., 2016 and Isola et al., 2011) could be relevant since the recalled image might reflect the

encoded image still present in the episodic memory. (need to add info about different representations for other types of information and how image memorability can be used and make two paragraphs)

2.3 Classification pipelines

The retrieval process in Wang et al., 2020b is tackled as a classification task (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.

Each encoding and recall histogram was fetched as a one channel input image by 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. 2020b varies among individuals, as those who actively moved their eyes during mental imagery performed better. 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.

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.

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. 2020b, 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.

2.4 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. 2020b addresses this issue by mapping the recall histograms to 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. 2020b 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. 2020b. 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.

2.5 Virtual Reality as an environment

An abundance of involuntary eye movements were observed when experiments were based on the looking-at-nothing paradigm (usually an empty screen or a white board) 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). Scholz et al. 2017 suggests that involuntary eye movements can be potentially driven by covert attention. 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 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 is presented during encoding.

So far the lack of reference frame proved to be problematic in previous studies including Wang et al. 2020b. 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. Asking participants to do an effort and visualise an upscaled version is not unreasonable since it is common practice in fMRI or EEG based studies to impose motor restrictions on participants. A reference frame can also contribute as implicit feedback for the participants during recall time and minimise the accumulation of errors over time.

A decreased signal-to-noise ratio would be hard to detect. 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 for this situation would also be avoiding distractions. As Wnag et al 2020b explains, " 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".

The minimisation of blinks might be hard in an as close to pitch black as possible 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.

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.

2.6 Conclusion

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

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

- can we improve on current state of the art?

RQ2 Can the impact of the scaling and translation factor in the recall eye movements be minimised using a VR environment and other classification pipelines or other mapping pipelines?

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

RQ4 Can longer recall time result in different histograms for different pictures but with similar layouts, enhancing classification accuracy?

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.

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.

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.

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.

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.

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.

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.

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.

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.

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, 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, 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`