ADVANCING THE USE OF EYE-TRACKING AND PUPILLOMETRIC DATA IN COMPLEX ENVIRONMENTS

EDITED BY: Russell A. Cohen Hoffing, Steven Matthew Thurman, Jonathan Touryan, Julien Epps, Josef Faller and Paul Sajda

PUBLISHED IN: Frontiers in Psychology and Frontiers in Neuroscience

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-015-2
DOI 10.3389/978-2-88976-015-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ADVANCING THE USE OF EYE-TRACKING AND PUPILLOMETRIC DATA IN COMPLEX ENVIRONMENTS

Topic Editors:
Russell A. Cohen Hoffing, CCDC Army Research Laboratory, Human Research and Engineering, US Army Research Laboratory, United States
Steven Matthew Thurman, CCDC Army Research Laboratory, Human Research and Engineering, US Army Research Laboratory, United States
Jonathan Touryan, United States Army Research Laboratory, United States
Julien Epps, University of New South Wales, Australia
Josef Faller, Columbia University, United States
Paul Sajda, Columbia University, United States

Citation: Hoffing, R. A. C., Thurman, S. M., Touryan, J., Epps, J., Faller, J., Sajda, P., eds. (2022). Advancing the use of Eye-Tracking and Pupillometric Data in Complex Environments. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-015-2

Table of Contents

frontiers in Psychology

ORIGINAL RESEARCH
published: 01 February 2021
doi: 10.3389/fpsyg.2021.604522

A Recurrent Neural Network for Attenuating Non-cognitive Components of Pupil Dynamics

Sharath Koorathota[1,2]*, *Kaveri Thakoor*[1], *Linbi Hong*[1], *Yaoli Mao*[3], *Patrick Adelman*[2] and *Paul Sajda*[1]

[1] Department of Biomedical Engineering, Columbia University, New York, NY, United States, [2] Fovea Inc., New York, NY, United States, [3] Department of Cognitive Science, Columbia University, New York, NY, United States

There is increasing interest in how the pupil dynamics of the eye reflect underlying cognitive processes and brain states. Problematic, however, is that pupil changes can be due to non-cognitive factors, for example luminance changes in the environment, accommodation and movement. In this paper we consider how by modeling the response of the pupil in real-world environments we can capture the non-cognitive related changes and remove these to extract a residual signal which is a better index of cognition and performance. Specifically, we utilize sequence measures such as fixation position, duration, saccades, and blink-related information as inputs to a deep recurrent neural network (RNN) model for predicting subsequent pupil diameter. We build and evaluate the model for a task where subjects are watching educational videos and subsequently asked questions based on the content. Compared to commonly-used models for this task, the RNN had the lowest errors rates in predicting subsequent pupil dilation given sequence data. Most importantly was how the model output related to subjects' cognitive performance as assessed by a post-viewing test. Consistent with our hypothesis that the model captures non-cognitive pupil dynamics, we found (1) the model's root-mean square error was less for lower performing subjects than for those having better performance on the post-viewing test, (2) the residuals of the RNN (LSTM) model had the highest correlation with subject post-viewing test scores and (3) the residuals had the highest discriminability (assessed via area under the ROC curve, AUC) for classifying high and low test performers, compared to the true pupil size or the RNN model predictions. This suggests that deep learning sequence models may be good for separating components of pupil responses that are linked to luminance and accommodation from those that are linked to cognition and arousal.

Keywords: recurrent neural network, pupil diameter, eye tracking, video viewing, pupil response

OPEN ACCESS

Edited by:
Guillaume Chanel,
Université de Genève, Switzerland

Reviewed by:
Juan Sebastian Olier,
Tilburg University, Netherlands
Walter Gerbino,
University of Trieste, Italy

***Correspondence:**
Sharath Koorathota
sharath.k@columbia.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *16 September 2020*
Accepted: *04 January 2021*
Published: *01 February 2021*

Citation:
Koorathota S, Thakoor K, Hong L, Mao Y, Adelman P and Sajda P (2021) A Recurrent Neural Network for Attenuating Non-cognitive Components of Pupil Dynamics. Front. Psychol. 12:604522. doi: 10.3389/fpsyg.2021.604522

1. INTRODUCTION

1.1. Pupillary Response

Physiological measures during cognitive processing have been extensively studied with pupillary dilation, in particular, having been explored as an index of learning, cognitive load, attention and memory (Sibley et al., 2011; Wang, 2011; Fridman et al., 2018). Dilation is generally understood to be mediated by increased sympathetic activity or inhibition of the parasympathetic

response (Karatekin, 2007) and reflected by activity in the brain's locus coeruleus-norepinephrine system (LC-NE), which controls physiological arousal and attention. LC-NE activity has been correlated with subjective task difficulty, cognitive effort, and neural gain (Eckstein et al., 2017). Mechanistically, the responsiveness of the pupil is driven by antagonistic actions of the iris dilator and sphincter muscles (Joos and Melson, 2012). Specific cognitive influences include pupil dilation in response to error in risk prediction and decision making (de Gee et al., 2014; Buettner et al., 2018), to emotional arousal (Hess, 1972), and in the presence of a known visual target (Privitera et al., 2010). In addition, the pupil has been shown to dilate to increased processing load in language tasks (Wang, 2011).

Pupil dilation is also important for regulating light entering the eye (Winn et al., 1994) and thus measures of cognitive processes linked to the pupil are confounded by: (1) the natural dilation changes due to luminance, (2) the photometric measure of light entering the eye, or (3) accommodation, the process by which the eye keeps focus on an object across varying distances. It is established that the pupil constricts with increasing luminance (Raiturkar et al., 2016), as the former is modulated by the pretectal nucleus. In fact, multiple studies have shown that luminance conditions take priority over cognitive demands in pupil diameter changes, across task difficulty and modality (Xu et al., 2011; Kun et al., 2012; Peysakhovich et al., 2015). Accommodation also effects pupil diameter to a lesser extent and appears to be limited as a driver in younger populations (Mathur et al., 2014).

1.2. Learning and Eye Tracking

In addition to pupillary response reflecting cognitive processing, past work has examined how other eye movements, such as fixations, can be indicators of cognitive processing when viewing educational content. Eye movements are more variable and less restricted by content boundaries in a younger audience while viewing Sesame Street, and video comprehension increases with age (Kirkorian et al., 2012). As visual and auditory saliency has strong direct impacts on visual exploration (which is captured by eye movement) and therefore indirect impacts on learning (Coutrot et al., 2014), eye movement information can be used to predict subjects' attention to viewing content.

The use of eye tracking data to help understand how students process content derived from different modalities has been employed to study how attention on PowerPoint slides changes with or without relevant narration (Slykhuis et al., 2005). Furthermore, viewing behavior has been used to assist in prediction of learning styles, using post-viewing assessments and viewing ratios (Cao and Nishihara, 2012) and, more recently, gaze behaviors such as fixations have been shown to vary with perceived relevance and presentation modalities of instructional content (Wiedbusch and Azevedo, 2020).

Simple eye tracking models have been employed to predict attention using measures such as total fixation duration (Xu et al., 2008). In our case, we seek to model how the input space predicts pupil dilation, using fixational and pupil features from eye tracking data along with contextual features from instructional video. While pupil dilation is most strongly affected by luminance-driven changes, recent work has yielded encouraging results in using pupil diameter to track lapses in attention (van den Brink et al., 2016), cognitive load (Wang, 2011) and as an index of learning (Sibley et al., 2011). One possible approach to distinguish between attention and luminance-driven effects is through comparison of model accuracy between above- and below-average performers in learning tasks. We hypothesize that in such a comparison, pupil diameter will be more variable and thus harder to predict in above-average performers, who may be more driven by pupil-linked arousal fluctuations.

1.3. Modeling Eye Tracking Data

To detect eye tracking events of interest, random forest models have previously been employed to detect fixations, saccades, and post-saccadic oscillations, yielding close-to-human level annotations (Buettner et al., 2018). Visual attention modeling has utilized video-level features, mapping these features to spatial and temporal saliency maps (Fang et al., 2017) in order to model gaze preferences. Bayesian networks and hidden Markov models have been used to learn patterns in eye movements to recognize facial expressions (Bagci et al., 2004; Datcu and Rothkrantz, 2004). Recent work has also analyzed still video frames through convolutional neural networks to analyze gaze data with the purpose of classifying groups (Dalrymple et al., 2019). However, sequences of fixations over areas of interest may also be useful in distinguishing individuals and groups (Çöltekin et al., 2010). In general, linear models, including those that employ regularized regression (ridge and lasso) (Papoutsaki et al., 2016) are simple and typically less likely to overfit the data. Non-linear models, including recurrent neural networks (RNNs) are interesting to consider as an alternative to linear models. For example, though RNNs are more complex and typically have more parameters then their linear counterparts, they can learn state sequence information over multiple timescales and feature dimensions. The long short-term memory model (LSTM) is a form of recurrent neural network that learns parameters over large amounts of sequence data efficiently (Hochreiter and Schmidhuber, 1997). LSTMs are used in language modeling, for example, as they are particularly suited to sequence data, and have been shown to outperform traditional deep learning network architectures (Sundermeyer et al., 2012; Koorathota et al., 2020). Because of this, the use of a sequence model such as an LSTM is a natural next step in analyzing gaze sequences.

1.4. The Present Study

The primary aim of the this study was to assess the prediction of pupil diameter in groups of participants whose performance varied on post-viewing assessments of educational content. We hypothesize that, due to the viewing dynamics, the realistic content, and the fact that information conveyed in the video is sparse compared to the length of the videos, a model that predicts pupil dynamics will tend to learn non-cognitive components, e.g., dynamics due to luminance changes, motion, accommodation. In this case we expect the residuals of the pupil dynamics under the model, i.e., those dynamics which are not predictable by the model, to be more informative of cognitive performance.

Toward that end, we initially compared accuracy of linear, non-linear, and RNNs when predicting pupil diameter. We further varied the type of input features we used as input to our models, to parse the usefulness of various eye movements and events when predicting pupil diameter. We then correlated the residuals from the most accurate models with performance on the post-viewing assessments to understand how accuracy of prediction varies across performers. We found that, compared to other models, the RNN (LSTM) (1) had root-mean square error (RMSE) that was less for lower performing subjects than for those having better performance on the post-viewing test, (2) the residuals of the model had the highest correlation with subject post-viewing test scores and (3) the residuals had the highest discriminability (assessed via area under the ROC curve, AUC) for classifying high and low test performers.

2. METHODS SUMMARY

2.1. Study Summary

61 healthy subjects (47 female, ages 18–35 with a mean of 25) participated in this study. Informed consent was obtained from all volunteers and the Columbia University Institutional Review Board approved all experiments. Participants were randomly assigned into three modality conditions to watch three 5-min-long lecture videos, with their eye movements recorded. After each video, they were instructed to answer a set of 7 multiple-choice questions, with a single correct answer, assessing comprehension of the video content just shown.

The lecture videos consisted of slides with images and bullet-point lists, presented by a professor in an academic classroom setting. Videos were produced to closely mimic the type of lecture students were likely to encounter in a real-life college-level academic setting as well as to provide sufficient context so that no subject-specific familiarity and expertise with the topic is required to answer the questions. The specific selection criteria for the lectures were as follows:

1. They had to be complex in content and be on topics that the participants were unlikely to be very familiar with but were also likely to find interesting,
2. They had to have visuo-spatial content that would allow for both images and a diverse set of gestures.

We chose the following three topics: the history of tarmac road paving, the use of perspectives in drawing, and the history of bicycles (**Figure 1A**). Additionally, speaker style and movement, as well as video editing techniques (cuts, edits, graphics, and sound effects) were also controlled in the video production using pre-specified scripts.

Of the questions assessing comprehension, 6 were slide-specific, in that the information used to answer each question was contained in one slide, and the remaining question required information across the presentation. The validity of questions were tested in a pilot study with 7 additional subjects so that ambiguous or unclear wording was clarified and items too difficult or easy were revised to have the proper discriminability to evaluate understanding.

The three modality conditions (i.e., video types) were produced with the same audio content but different types of visual content, including single (full-screen slides), dual (slides and audio lecture), and full (professor with upper body view visible on the lecture video, with slides present) versions. Using a between-subject design, each subject was shown the same modality version for all three topics—controlling for luminance across viewing sessions. The topics were always presented in the same order: history of road paving, visual perspectives, and history of bicycles.

2.2. Eye Movements and Pupil Dilation

Eye tracking was performed with an Eyelink 1000 in Tower Mount, at a sampling rate of 1 kHz. Eye tracking data contained X and Y coordinates of each fixation (pixels), fixation duration (ms), pupil diameter (μm), saccades, blinks and associated timestamps (**Figure 1B**).

Subjects were instructed to watch videos presented on a 30-inch screen from 40 inches away without wearing glasses. The study was conducted in a Faraday's cage with low-light, sound-proof conditions that remained constant during video watching. Before each of the three videos, the eye tracker was calibrated for each recruited subject. In the calibration procedure, subjects were asked to focus their gaze on nine points presented consecutively at specific positions across the diagonals and centers of the side edges of the display screen. Moreover, subjects were instructed not to move their heads and to pay attention to the lecture content presented on the screen throughout video watching.

For each subject, we filtered for fixations out of the video frame boundary and systematic drifts. 3 participants were found to spend a non-negligible amount of time (>6%) blinking or fixating outside of the center rectangle video frame boundary and were excluded as outliers, leaving a total of 58 subjects for further analysis.

Classifications of eye events, including fixations, saccades, and blinks were exported from the SR Research software, which uses video-oculography based classification algorithms and pupil diameter calculations.

2.3. Problem Types

The prediction problems or inputs varied across two dimensions: (1) the amount of time, relative to the input, used in the generation of the output label and (2) the types of input features used for predictions.

We utilized five categories of input features for the models:

- Fixations: positions, durations, start times, and respective differences from fixation to fixation,
- Pupil diameter: per fixation,
- Areas of interest: a mapping of sequence of AOI to 50-dimensional embeddings learned during training process,
- Saccades: saccade-related positions, durations, start times and respective differences,
- Blinks: blink times and differences.

We investigated the effect of various combinations of the types of inputs: {fixations, fixations + pupil diameter, fixations + saccades + blinks fixations + pupil diameter + saccades + blinks}.

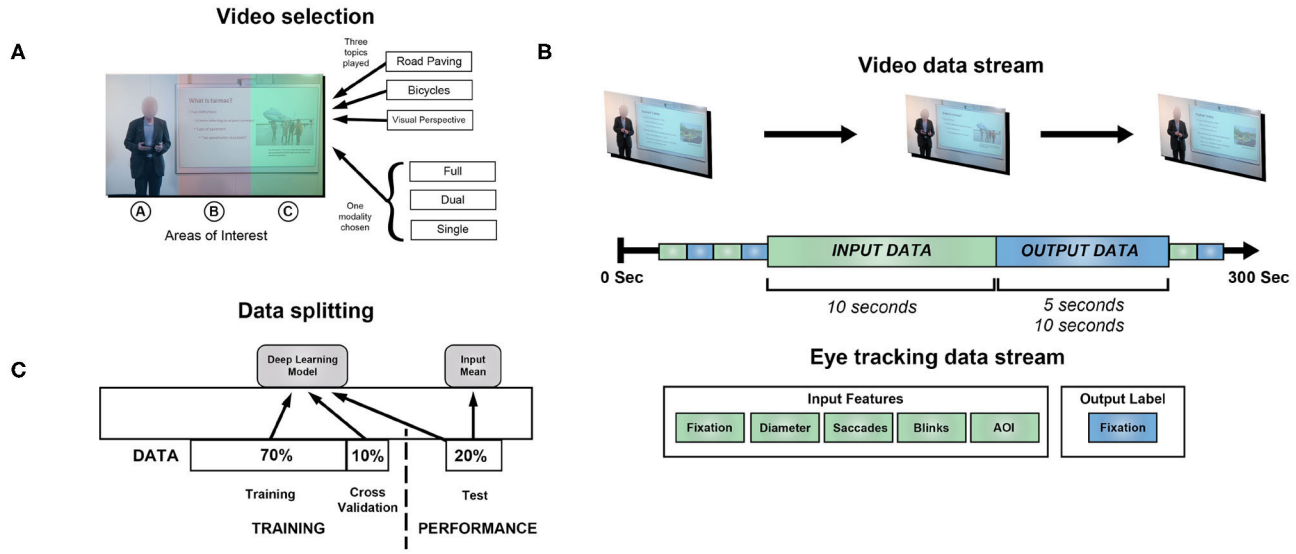


FIGURE 1 | Overview of methods. **(A)** Participants viewed videos on three topics while eye tracking data was collected: visual perspective, bicycles, and road paving. They were randomized to one of three modalities: full, dual or single. Colorized in the image are the areas of interest (AOI) used in training several of the models we consider. Subjects did not see this colorization nor were they explicitly aware of the AOIs. **(B)** The video spanned 300 s, while the eye tracking data was split into 10 or 15 s blocks over the course of viewing. The first 10 s was used as input data toward the model, with various types of features. **(C)** The blocks of eye tracking data were split into training, validation and test data for model fitting and testing. Predictions were derived from the model frameworks, e.g., the deep learning model, and the mean of the input pupil diameter, for a naive estimate.

Because eye tracking data can be sourced from web cameras, infrared devices, or human annotations, each with varying level of accuracy for labeling eye movements and events, our aim was to assess the minimal amount of data that yields accurate predictions of pupil diameter. We were not able to find similar iterative approaches to predicting pupil diameter using different types of input data and hypothesized historical fixation and input pupil diameter to be the best predictors of future pupil diameter.

In addition, for baseline reference, we report the error rates in models that are most commonly used toward prediction of eye tracking data:

- Linear regression: simple linear fit of input features,
- Regularized regression: linear regression with penalization of large weights through L1 (Lasso) and L2 (Ridge) norms,
- Decision-tree based: ensemble learning methods relying on majority vote by weak models (gradient boosting) or mean of trees (random forest),
- Input mean: a naive estimate of the mean pupil diameter in the input.

Hyperparameters for the reference models were selected from default recommendations from scikit-learn, a popular machine learning framework in Python (Pedregosa et al., 2011).

2.4. Data Aggregation

Because this study was supplementary to a larger one focusing on the effects of gestures on learning, we were presented the option to use data from single or multiple modalities. The justification for using all available modalities for prediction of pupil diameter was twofold: to allow for a large enough amount of data to utilize deep learning models that we predicted would perform well, and to increase the robustness of prediction of pupil diameter under different modalities of learning. Because, in a natural learning environment, students may be presented with video and audio but may not necessarily attend to it (Chen and Wu, 2015), this dataset provided a unique opportunity to predict pupil diameter and assess model accuracy under mixed modalities.

As a first step for analysis, eye tracking streams were split into 15-s blocks, across all participants, modalities and topics, and randomized. The first 10 s in each block were used to sequence input data, while pupil diameters in fixations in the succeeding 5 s of the block were averaged to yield the associated output label. In another method of analysis, eye tracking streams were split into 20-s blocks, with features collected over the first 10 s as input and the succeeding 10-s fixation pupil diameter as output.

Subsequent analyses are reported from the best-performing model using 10 s of input to predict 5 s of output. We made this selection in order to maximize the number of samples and use typical output durations studied in past eye fixation work (Just and Carpenter, 1976).

Due to this method of data aggregation, the number of fixations, saccades and blinks varied across and within participants. Thus, the input region required feature-specific, mean padding to the maximum length of fixations. The output was always a single-dimensional, average, fixation pupil diameter gathered from the output region. Thus, the deep learning models can be thought of as regression problems utilizing a non-linear framework.

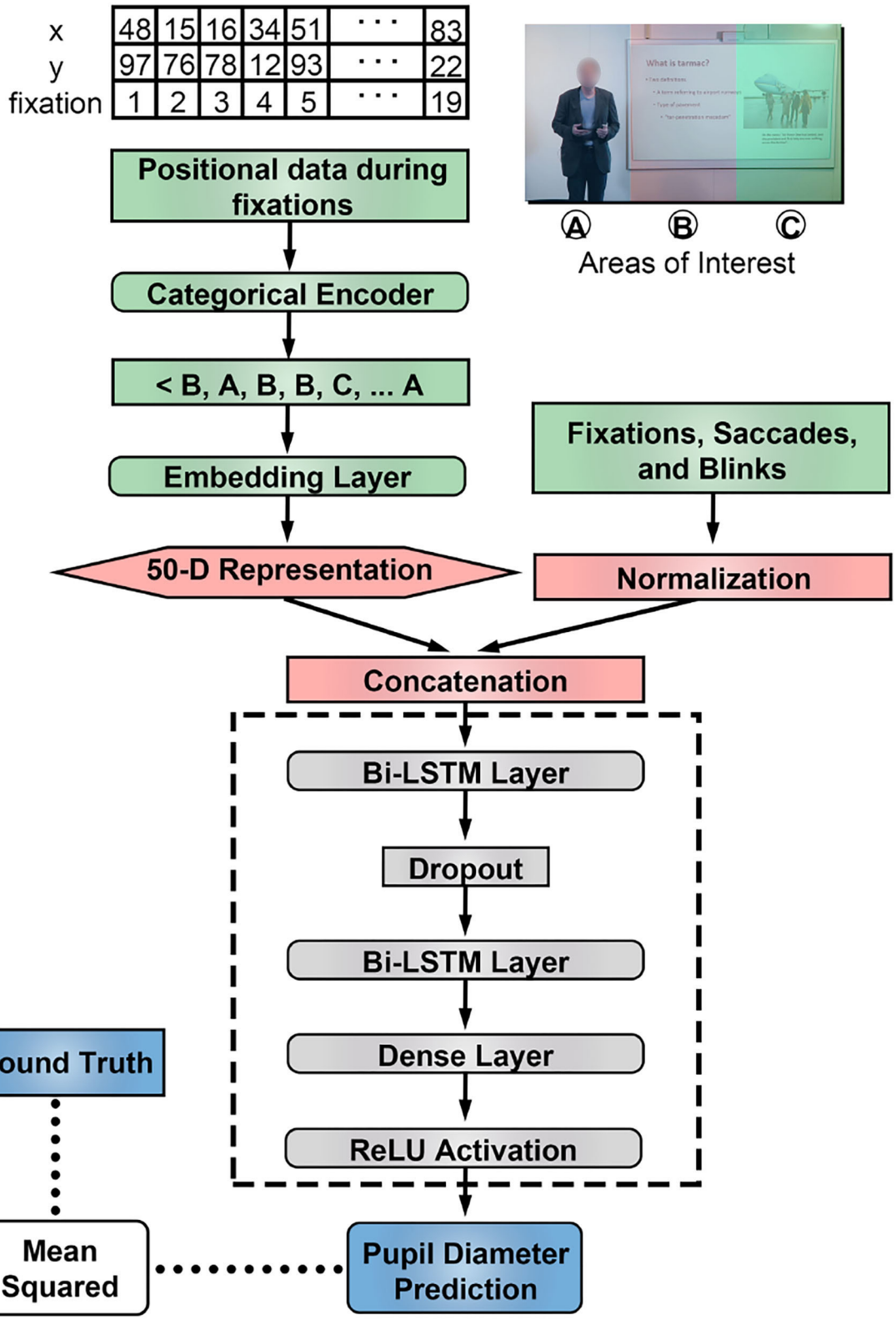


FIGURE 2 | LSTM architecture included two bidirectional layers as the core component. Numerical features were normalized and the areas of interest were embedded to a higher dimensional vector trained using the training samples. Embeddings are trained using categorical representations of fixation data.

TABLE 1 | Network hyperparameters.

Param	Value
Epochs	5,000
Early stopping	500 epochs (loss)
Optimizer	Adam
Learning rate	0.0001
Training split	70%
Validation split	10%
Test split	20%

2.5. AOI Embeddings

We defined three, distinct, areas of interest (AOI) in the full video type, across all topics, corresponding to the instructor, text in slides and images in slides. Other types (dual and single) contained only text and image AOI.

We mapped X and Y coordinates from fixations in input regions to AOI. This allowed us to generate a sequence of AOIs for fixations during a specified input region, which we used to train 50-dimensional embeddings during the training process (**Figure 2**). We hypothesized that this process will achieve a similar goal as in natural language applications of capturing context of categorical information with respect to other input features (Melamud et al., 2016).

2.6. Network Architecture

We used a bi-directional LSTM network to model eye tracking input (**Figure 2**). For each problem type, data was split into training, validation and test samples (**Figure 1C**). The network was trained and validated on estimates of pupil diameter and assessed through mean-squared error using the Adam optimizer (**Table 1**). Each LSTM layer used recurrent dropout to prevent overfitting.

We compared our network's results to the mean pupil diameter as calculated from the fixations in the input region and with other, reported linear and non-linear approaches. In addition, we compared our LSTM network results to a gated recurrent unit (GRU), an RNN variant (Chung et al., 2014), with the same network hyperparameters and without recurrent dropout. Neural models were implemented in Tensorflow 2.2 on Google Cloud and trained using a NVIDIA Tesla K80 GPU.

2.7. Data Analysis

We hypothesized that predictability of pupil diameter would vary across four dimensions: (1) as a ratio of input time (used in the aggregation of input features) and output time (used to calculate for ground truth pupil diameter), (2) use of different physiological measures in the network, (3) addition of AOI embeddings in the neural network, and (4) participant performance on the post-viewing assessments.

To test the hypothesis that predictability varied across the four dimensions, we first split the data into 15-s blocks. We designed a baseline comparison through averaging the pupil diameter across fixations during the first 10 s of each block. This served as the naive, input mean, estimate.

Using the first 10 s in each block to aggregate input features, we randomly separated the data into training, validation and test sets, calculated the RMSE to study the prediction errors in the test set. We repeated the process a total of 10 times (i.e., runs) for each problem type and using different input features in the best-performing model to account for variability of accuracy due to the training and test separation of the data. Furthermore, we repeated the process above after separately splitting the data into 20-s blocks first, predicting 10 s of output. We summarize the reported RMSE measure

$$RMSE_{M,I,O} = \frac{1}{l}\sum_{t=1}^{l}\left(\sqrt{\frac{1}{n}\sum_{s=1}^{n}(\hat{y}_s - y_s)^2}\right),$$

where $\hat{y}_s$ is the predicted pupil diameter, y_s is the ground truth, output pupil diameter, n is the number of training samples, and l is the number of random, training, validation and test splits RMSE was averaged over (always 10). This value was calculated for each model type, M, for different sets of input features, I, and output period length O over which fixation pupil diameters were averaged.

The aggregation and split of the data led to reusing the same 15- or 20-s blocks across the 10 runs. These were treated as independent samples, regardless of the video type, condition or participant they originated from.

2.8. Participant Performance

To study model accuracy in groups with different levels of cognitive effort, we split the test blocks by mean performance on post-viewing assessments (i.e., into "Greater Than Mean" and "Lesser Than Mean" bins). We report model results separately for these groups, using a Mann–Whitney U-test for significant, mean differences in model errors.

Using residuals from the most accurate model, we report Spearman correlation coefficients in the test samples between the ground truth pupil diameter, the estimate from the model, residuals (ground truth minus model estimate) and performance. To assess the predictive accuracy directly, we designed a simple binary classification task using the ground truth pupil diameter, model estimate and residuals to classify participants as belonging to the lowest or highest tertile group by performance. We used an ROC analysis, which consists of a plot of the sensitivity and 1-specificity pairs that are produced as a single decision threshold is moved from the lowest (i.e., all participants classified in the lowest tertile) to the highest (i.e., all participants classified in the highest tertile) possible value (Fawcett, 2006). The area under the ROC curve (AUC) corresponds to the probability that a randomly selected participant will have been assessed by the measure (e.g., residuals) as performing better than a randomly selected participant, and varies from 0.5 (i.e., accuracy is not improved over chance) to 1.00 (i.e., perfect accuracy).

Thus, we used group-level RMSE differences to quantify how model accuracy varies with levels of cognitive effort and residuals to understand the relation between the accuracy of model predictions and participant performance.

TABLE 2 | Prediction errors (RMSE) for linear, regularized linear, decision-tree based, and RNN (GRU, LSTM) model types.

Inputs	Model type	RMSE
Fixation + Diameter + Saccades + Blinks	Linear regression	>5000
	Ridge regression	332.72 (5.11)
	Gradient boosting	319.14 (12.05)
	Input mean	312.93 (13.32)
	GRU	300.91 (20.64)
	Lasso regression	295.10 (9.18)
	Random forest	292.79 (12.26)
	LSTM	285.65 (9.69)
Fixation + Diameter + Saccades + Blinks + Embeddings	Linear regression	>5,000
	Ridge regression	332.35 (13.38)
	Gradient boosting	323.56 (13.34)
	Input mean	306.45 (11.13)
	Lasso regression	304.34 (9.61)
	Random forest	298.05 (11.45)
	GRU	288.38 (12.30)
	LSTM	249.87 (8.65)

Inputs from 10 s of each block was used to predict 5 s of subsequent, average, fixation pupil diameter. Input mean refers to the naive estimate of using the mean pupil diameter in the input data as a prediction of the pupil diameter.
Mean (SD).

3. RESULTS

A total of 2,379, 20-s blocks and 3,249, 15-s blocks were analyzed, with an average pupil diameter of 2126.42 μm (SD = 916.04 μm) and 2134.33 μm (SD = 934.20 μm) respectively.

3.1. Model Comparisons

We first report the mean error metrics, averaged over 10 runs, for each model type in **Table 2**. The use of embeddings improves the model accuracy only for the LSTM, which also outperforms the other model types we tested in average RMSE. For the remaining results, we utilized the best performing model, the LSTM.

3.2. Input Feature Comparisons

We report the aggregate accuracy, in terms of RMSE with respect to ground truth pupil diameter, of the LSTM models and the input mean (**Table 3**). When pupil diameter was used as an input, RMSE was significantly lower than the input mean model (312.93 μm). The best performing model used only fixation and pupil diameter measures as input, with 10 s of input predicting mean pupil diameter for 10 s of output. This had a mean RMSE of 252.97 μm.

Generally, when pupil diameter was used as an input, accuracy significantly improved as output length increased from 5 to 10 s.

3.3. Addition of Embeddings

Next, we report the change in RMSE as a result of adding the AOI embeddings (**Table 3**). When using pupil diameter as an input, adding AOI embeddings significantly reduced the RMSE. In these cases, the drop in RMSE was significantly more than 35 μm, with a more pronounced effect when predicting output pupil diameter in 5 s. The effect of AOI was less pronounced when predicting pupil diameter averaged over the longer time span of 10 s, indicated by less reduced RMSE and non-significant reductions even in the condition utilizing the full set of input features (−9.68 μm, $p > 0.05$). Note, subsequent analyses is reported only for the 5 s output condition.

3.4. Performance Differences

The average, post-lecture, performance on the assessment was determined to be 59% across participants, video types and conditions. Thus, we report the accuracy of the LSTM and input mean models in participants who performed greater or lesser than this mean.

In all cases, model accuracy was relatively better in participants who scored below the mean (**Table 4**). In the best-performing case (using fixation and pupil diameter as input), the RMSE, on average, decreased by 31.13 μm ($p < 0.01$) when using the same model for below-average compared to above-average performers.

The input features whose associated accuracy resulted in the greatest difference between groups, surprisingly, was the input mean pupil diameter, showing a significant difference of 64.53 μm ($p < 0.01$) between below- and above-average performers. All other frameworks, using different input features, experienced better prediction in the below-average performers ($p < 0.01$).

We found a similar pattern of reduction as in the case of aggregate analysis (**Table 3**) in RMSE after adding in AOI embeddings for both above- and below-average performers.

We also computed the correlation between ground truth, estimated, and residual (ground truth minus estimate) pupil diameter with participant performance (**Figure 3A**). Performance correlated significantly (at the 0.01 significance level) with the residuals from the LSTM model ($r = 0.33$), but not the true pupil diameter ($r = 0.24$) or the LSTM estimate ($r = 0.21$) at the 0.05 level. A Fisher Z-test showed that the difference between the correlations derived from the residuals and true pupil diameter were not significantly different at the 0.05 level ($z = 0.66$). We plot the distributions, by modality, of the true pupil diameter (mean ± SD): 2285.57 ± 1237.60 μm full, 2024.77 ± 677.94 μm dual, 1981.92 ± 758.19 μm single; LSTM estimate: 2252.86 ± 1031.46 μm full, 2018.38 ± 544.00 μm dual, 1979.03 ± 609.66 μm single; and residual: 32.72 ± 357.61 μm full, 6.39 ± 271.39 μm dual, 2.89 ± 315.49 μm single in the test samples (**Figure 3B**). Interesting to note is that the residuals of the model are more invariant to the variations in modality type, then the actual pupil measures or the models predictions. This is likely to reflect variation in non-cognitive measures across modality that are captured by the model and are attenuated in the residuals.

As a further analysis, we computed the separation between performance group classes (i.e., highest and lowest tertile of mean post-viewing test scores) using AUC measures (see **Figure 4**). AUC was largest for the model residuals compared to the model prediction and true pupil diameter measurements ($AUC_{residuals} = 0.74$, $AUC_{LSTM} = 0.63$, $AUC_{pupil} = 0.65$). To construct a null for significance testing, we performed 10,000 permutations of class labels and found residuals-derived AUC ($p < 0.01$) and true pupil diameter-derived AUC ($p = 0.05$) were significantly greater

TABLE 3 | RMSE test accuracy for given set of input features using the LSTM framework, including a simple comparison using the mean pupil diameter across fixations.

Input features	10 s input predicting 5 s output			10 s input predicting 10 s output		
	RMSE	$\Delta RMSE_{AOI}$	*n*	RMSE	$\Delta RMSE_{AOI}$	*n*
Fixation	711.77*** (14.06)	7.7	650	723.45*** (32.06)	−15.19	476
Fixation + Saccades + Blinks	652.74*** (20.66)	−13.71	650	661.81*** (33.00)	−9.05	476
Mean Input Diameter	312.93 (13.32)	-	650	302.98 (14.19)	-	476
Fixation + Diameter + Saccades + Blinks	285.65** (9.69)	−35.78***	650	266.27*** (12.65)	−9.68	476
Fixation + Diameter	270.71*** (10.74)	−35.12***	650	252.97*** (9.35)	−16.91***	476

Metrics are reported as mean (SD) and were averaged across 10 random test splits. Differences between the input mean to other models were assessed using Mann–Whitney U-Test. Deltas indicate differences in test accuracy measures after adding AOI embeddings to models, assessed using the Mann–Whitney U-test.
** ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.*

TABLE 4 | Mean (SD) accuracy differences after splitting data into above- and below- average (0.59) performers on the post-viewing assessments using the LSTM framework.

Input features	Greater Than Mean				Lesser Than Mean		
	RMSE	$\Delta RMSE_{AOI}$	*n*	$\Delta RMSE$	RMSE	$\Delta RMSE_{AOI}$	*n*
Fixation	736.68*** (34.78)	30.70	281	**	692.04*** (22.98)	−8.43	368
Fixation + Saccades + Blinks	679.21*** (38.70)	−7.84	284	**	632.6*** (18.46)	−18.34*	366
Mean input diameter	347.48 (24.60)	-	284	**	282.95 (10.13)	-	366
Fixation + Diameter + Saccades + Blinks	296.26*** (14.00)	−29.52***	284	**	277.19*** (11.91)	−40.71***	366
Fixation + Diameter	288.21*** (16.02)	−34.23***	284	**	257.08*** (15.10)	−36.00***	365

Significance assessed using the Mann–Whitney U-test for mean differences in RMSE (across 10 random, test data splits or model runs) between sets of input features and mean input pupil diameter, and separately for delta scores after addition of AOI ($\Delta RMSE_{AOI}$) within groups. We also report differences in between the above- and below-average performers using various input features (Δ RMSE).
** ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.*

than chance while the model prediction-derived AUC was not. This provides further evidence that the residuals of the model are informative of cognitive performance.

4. DISCUSSION

Using viewing instructional video as a test case, we found that an LSTM recurrent network was able to indirectly disentangle luminance and cognitive processes that affect pupil dilation. The effect is indirect in that the LSTM appears to better model non-cognitive components of the pupil dynamics. For example we see higher RMSE for subjects performing better on the post-lecture assessments, while conversely, lower RMSE for those performing less well.

Since the model was trained just to predict pupil response and not cognitive effort, it is reasonable to assume most of the pupil dynamics will be attributable to non-cognitive factors given the information presented in the video is temporally sparse relative to the length of the video. Thus, under our assumptions that:

1. Higher performance in the post-viewing assessments correlates with increased cognitive performance or effort and
2. Cognitive effort is more difficult to model than lower-level drivers of pupil diameters like luminance,

we believe our sequence networks are modeling changes in the pupil dilation that reflect luminance changes, and thus model the confound that researchers often try to control for when studying attention through eye tracking data.

This finding is strengthened by the significance of correlation between LSTM residuals and performance. The LSTM thus may act as a filter to attenuate non-cognitive information in the pupil dynamics, with the residuals of the resulting signal reflecting cognitive components of the pupillary response. AUC measures followed similar trends, with a simple, binary classifier yielding better accuracy in separating performance groups using the residuals over the true pupil diameter and LSTM estimates. We recommend future paradigms use more extensive assessments to improve statistical power in related tasks.

4.1. Pupil Diameter Prediction

Under constant, 15.9 lux ambient illumination, pupil sizes for males and females aged 19 have been reported to vary around a mean of approximately 7,100 μm by 900 μm (one SD) (MacLachlan and Howland, 2002). Given this fact, even the simple, input mean is a reasonable predictor of pupil diameter during video viewing (**Table 3**). However, the best performing model (LSTM using fixation + diameter + saccades + blinks + AOI) provides a much more narrow estimate (235.59 μm) of pupil diameter across all participants. We attribute this increased accuracy to the non-linear learning capability of LSTMs, which appear to successfully learn relationships between the input features and, especially using the contextual information stored

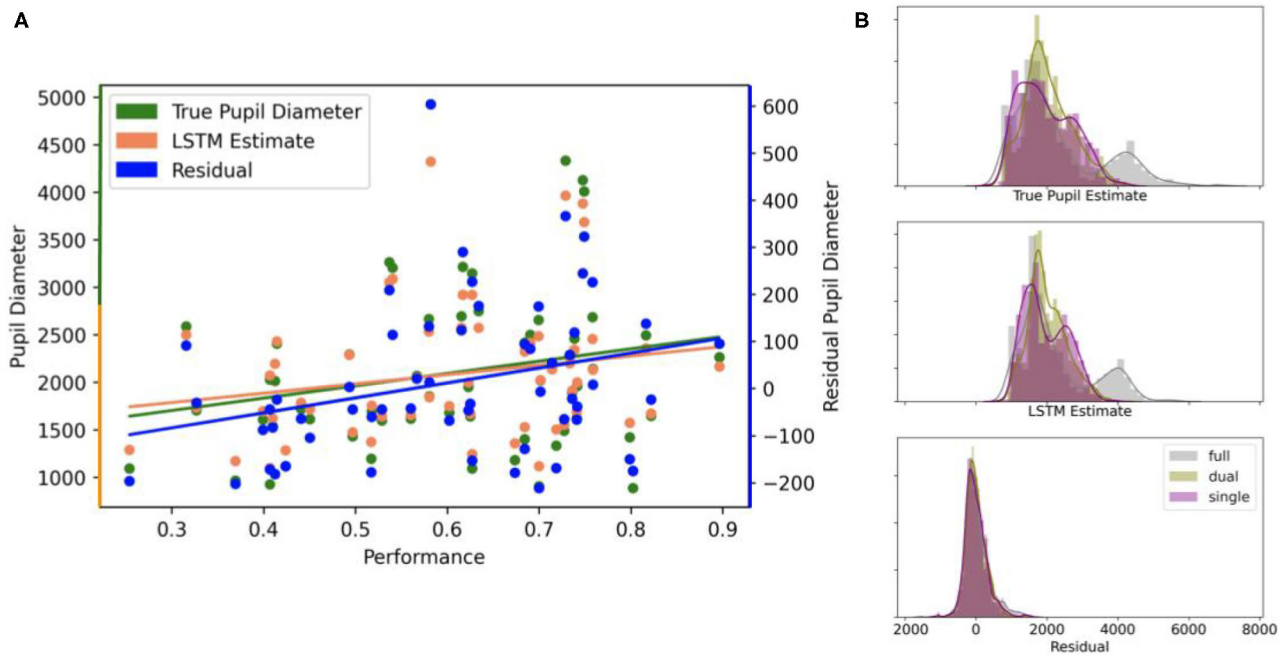


FIGURE 3 | (A) Correlational relationships in test samples, averaged by participant, between the true (ground truth) pupil diameter, the estimate from the LSTM model, and the residual (ground truth minus model estimate) with participant performance. True pupil diameter (green) and the LSTM estimate (orange) values use the left axis scale, while residual values (blue) use the right axis scale. **(B)** Distributions of true pupil diameter, LSTM estimates, and residuals by video modality.

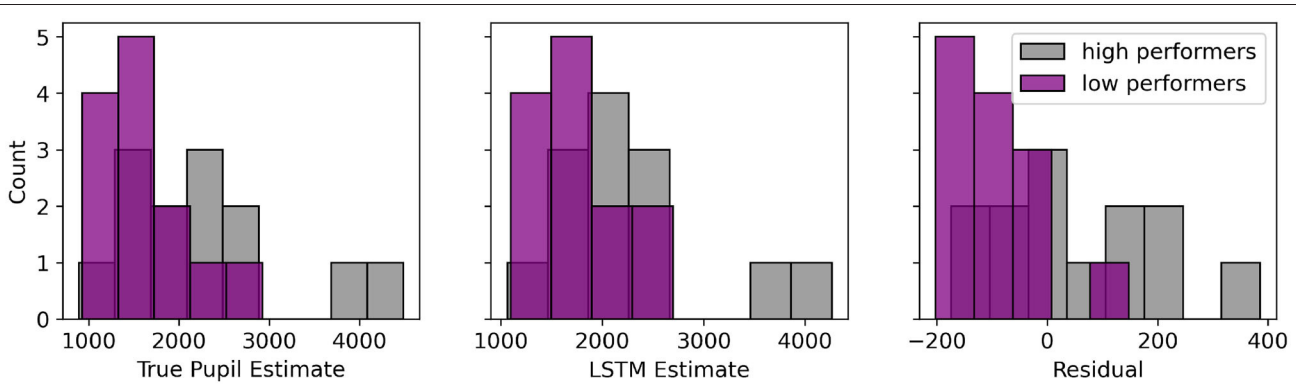


FIGURE 4 | Histograms of true pupil diameter, LSTM estimates and residuals of the bottom (purple, n = 13) and top (gray, n = 13) tertile of participant performance on post-viewing assessments. The AUC is calculated for each measure. ($AUC_{residuals}$ = 0.74, AUC_{LSTM} = 0.63, AUC_{pupil} = 0.65).

in AOI embeddings, predict pupil diameter with relatively low error in the test sets. While the GRU counterpart also had reduced prediction errors relative to the linear models, we note that the average RMSE was greater than the LSTM, and the variability in performance was larger. Furthermore, the GRU model performs worse, relative to the LSTM, when AOI embeddings are not used as input (**Table 2**). This may be due to the relatively increased control that the LSTM network architecture provides, which in this case may have improved the modeling of input eye events. In fact, this finding is consistent with existing literature showing RNN results vary with the complexity of sequences in a dataset (Chung et al., 2014).

Other non-linear models we evaluated for prediction of pupil diameter included random forests and gradient boosted regression trees. We hypothesized, due to the aforementioned benefits of non-linear models, that errors would be reduced for non-linear models compared to their linear counterparts. This was generally true, but the linear methods with regularization (i.e., Lasso and Ridge Regression) were similar in their error rates to non-linear methods.

We interpret the findings from reduced error rates using recurrent methods, relative to the naive, input mean estimate, to support the view that temporal memory is critical for accurate prediction of pupil diameter using eye tracking data.

This accurate prediction may provide more opportunities for human-computer interaction through inferring cognitive state (Medathati et al., 2020). While, our videos' intrinsic characteristics (e.g., luminance, hue) may be highly correlated with video AOIs and this may extend to correlation with pupil diameter for bottom-up processes that rely on stimuli saliency, we believe this extension complements the goal of our study. In fact, we train our models on data from multiple modalities for this reason precisely—because we believe that a video's intrinsic characteristics might be confounds for pupil dynamics and not assessment performance, and modeling approaches may work better for saliency-driven pupil changes and not cognitively-driven changes.

4.2. Improvement From AOIs

We fixed AOIs to be constant across videos, since we wanted to isolate regions most relevant to information processing in the given task. By controlling where and how information is presented in the videos, we attempted to study the effect of information presentation (e.g., through controlled text placements and instructor gestures) on pupil diameter. Our sequence model approach generally worked best when including not just eye tracking features but also context (via AOI embeddings). In all cases, adding AOI reduces RMSE—significantly in cases where pupil diameter is used as an input. Our findings indicate that pupil diameter, paired with fixational positions, provide a richer context of viewing patterns that allow accurate predictions of pupil diameter. We found a greater decrease in error when adding AOI embeddings as input predicting 5 s of average fixation pupil diameter. However, we believe this may be due to a floor effect since the difference yields RMSEs that are relative close in magnitude to the fixation + diameter input features from the 10 s output problem type.

While the information contained in embeddings is redundant with the fixation positions, we believe the categorical representation of continuous data (i.e., three AOIs from the large space of possible fixation coordinates) improved LSTM learning to yield lower error rates. In fact, architectures designed with characteristics of sparse data in mind during design tend to optimize faster and avoid local minima (Duchi et al., 2013).

4.3. Input Features

In our tested cases, we did not find significant improvements to our model after including saccade and blink sequences to fixation and pupil diameter inputs. We believe this may be because saccades and blinks are not related to pupil diameter in a task that requires focus such as in instructional video viewing. Despite a lack of human research related to our finding, we note animal research where microstimulation affected pupil dilation independently of saccades (Wang et al., 2012), highlighting the limited association of covert attention to pupil dilation. Because we partly sought a study of the minimum amount of eye tracking data required to accurately predict pupil diameter, our findings show that input features like saccades and blinks are not as critical as fixation and pupil diameter data when predicting future pupil diameter. We expect this finding to be helpful when focusing efforts for algorithms modeling pupillary mechanisms.

We note that our framework allows for prediction of other averages of eye tracking measures, such as fixation duration during the output region, blink rate, AOI-specific measures, etc. In addition, a framework such as ours allows for prediction of sequences of data—for example, fixation positions or pupil diameters. In fact, these types of problems mirror those faced in natural language processing, where deep learning, sequence models have performed relatively better than other linear or non-linear models for sequence outputs. Future work is required in applying this to viewing patterns.

4.4. Limitations

The primary limitation of our study is the lack of interpretability for the best-performing (LSTM) model, a common problem in deep learning studies. In this case, however, we attempted to solve the problem of not being able to understand the precise importance of input features by studying the effect of various models with modular inputs. We believe that this approach, paired with multiple runs of models to get average accuracy, addresses issues of interpretability and can be expanded upon in future work.

Additionally, we acknowledge that the LSTM model may be difficult to generalize to some training sequences. Our results on model accuracy, given modular inputs, allows some generalizability to sensors that are unable to extract pupil diameters or classification models unable to specify eye events such as saccades. However, a limitation of our approach is the lack of specificity of which LSTM hyperparameters or characteristics of eye events may be contributing to better accuracy of prediction. While our focus was on studying the effectiveness of RNNs in improving pupil prediction accuracy, and how student performance differences may be related to model accuracy, future work in this area should apply the same modularity within RNNs to further understand why deep learning models more effectively capture behavioral variations relative to their non-linear counterparts.

5. CONCLUSION

Our evaluation shows that, using AOI embeddings and fixation and pupil sequence history, a deep learning, sequence model predicts pupil diameter better than a naive mean-based estimate. Prediction is better for subjects who perform relatively poorly on post-lecture assessments, and model errors correlate with performance as a trend. This latter finding may indicate that those individuals were less engaged and thus had less expression of their cognition in their pupil dilation, allowing the model to capture luminance-influenced variations.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Columbia University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SK, LH, YM, KT, and PS conceived of the presented idea. SK developed the theory, performed the computations, and took the lead in writing the manuscript. PA provided data visualizations. LH and PS supervised the project. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

FUNDING

This work was supported by grants from the National Science Foundation (IIS-1513853 and IIS-1816363), the Army Research Laboratory Cooperative Agreement W911NF-10-2-0022 and a Vannevar Bush Faculty Fellowship from the US Department of Defense (N00014-20-1-2027).

REFERENCES

Bagci, A. M., Ansari, R., Khokhar, A., and Cetin, E. (2004). Eye tracking using Markov models. *Proc. Int. Conf. Pattern Recogn.* 3, 818–821. doi: 10.1109/ICPR.2004.1334654

Buettner, R., Sauer, S., Maier, C., and Eckhardt, A. (2018). "Real-time prediction of user performance based on pupillary assessment via eye-tracking," in *AIS Transactions on Human-Computer Interaction*, 26–60. doi: 10.17705/1thci.00103

Cao, J., and Nishihara, A. (2012). Understand learning style by eye tracking in slide video learning. *J. Educ. Multimedia Hypermedia* 21, 335–358.

Chen, C.-M., and Wu, C.-H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Comput. Educ*. 80, 108–121. doi: 10.1016/j.compedu.2014.08.015

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint* arXiv:1412.3555.

Çöltekin, A., Fabrikant, S. I., and Lacayo, M. (2010). Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *Int. J. Geogr. Inform. Sci.* 24, 1559–1575. doi: 10.1080/13658816.2010.511718

Coutrot, A., Guyader, N., Ionescu, G., and Caplier, A. (2014). Video viewing: Do auditory salient events capture visual attention? *Ann. Telecommun*. 69, 89–97. doi: 10.1007/s12243-012-0352-5

Dalrymple, K. A., Jiang, M., Zhao, Q., and Elison, J. T. (2019). Machine learning accurately classifies age of toddlers based on eye tracking. *Sci. Rep*. 9, 1–10. doi: 10.1038/s41598-019-42764-z

Datcu, D., and Rothkrantz, L. J. (2004). "Automatic recognition of facial expressions using bayesian belief networks," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (The Hague), 2209–2214.

de Gee, J. W., Knapen, T., and Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proc. Natl. Acad. Sci. U.S.A*. 111, E618-E625. doi: 10.1073/pnas.1317557111

Duchi, J., Jordan, M. I., and McMahan, B. (2013). "Estimation, optimization, and parallelism when data is sparse," in *Advances in Neural Information Processing Systems*, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Lake Tahoe, CA: Curran Associates), 2832–2840.

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., and Bunge, S. A. (2017). Beyond eye gaze: what else can eyetracking reveal about cognition and cognitive development? *Dev. Cogn. Neurosci*. 25, 69–91. doi: 10.1016/j.dcn.2016.11.001

Fang, Y., Zhang, C., Li, J., Lei, J., Da Silva, M. P., and Le Callet, P. (2017). Visual attention modeling for stereoscopic video: a benchmark and computational model. *IEEE Trans. Image Process*. 26, 4684–4696. doi: 10.1109/TIP.2017.2721112

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett*. 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Fridman, L., Reimer, B., Mehler, B., and Freeman, W. T. (2018). "Cognitive load estimation in the wild," in *Proceedings of the 2018 chi Conference on Human Factors in Computing Systems* (Montreal, QC), 1–9. doi: 10.1145/3173574.3174226

Hess, E. H. (1972). "Pupillometrics. a method of studying mental, emotional, and senosory prrocesses," in *Handbook of Psychophysiology*, 491–531.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Joos, K. M., and Melson, M. R. (2012). "Control of the pupil," in *Primer on the Autonomic Nervous System*, eds D. Robertson, I. Biaggioni, G. Burnstock, P. A. Low, and J. F. R. Paton (San Diego, CA: Elsevier), 239–242. doi: 10.1016/B978-0-12-386525-0.00049-4

Just, M. A., and Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cogn. Psychol.* 8, 441–480. doi: 10.1016/0010-0285(76)90015-3

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Dev. Rev*. 27, 283–348. doi: 10.1016/j.dr.2007.06.006

Kirkorian, H. L., Anderson, D. R., and Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child Dev*. 83, 497–507.

Koorathota, S. C., Thakoor, K., Adelman, P., Mao, Y., Liu, X., and Sajda, P. (2020). "Sequence models in eye tracking: predicting pupil diameter during learning," in *ACM Symposium on Eye Tracking Research and Applications* (virtual event), 1–3. doi: 10.1145/3379157.3391653

Kun, A. L., Palinko, O., and Razumenić, I. (2012). "Exploring the effects of size and luminance of visual targets on the pupillary light reflex," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Portsmouth), 183–186. doi: 10.1145/2390256.2390287

MacLachlan, C., and Howland, H. C. (2002). Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years. *Ophthal. Physiol. Opt*. 22, 175–182. doi: 10.1046/j.1475-1313.2002.00023.x

Mathur, A., Gehrmann, J., and Atchison, D. A. (2014). Influences of luminance and accommodation stimuli on pupil size and pupil center location. *Investig. Ophthalmol. Visual Sci*. 55, 2166–2172. doi: 10.1167/iovs.13-13492

Medathati, N. V. K., Desai, R., and Hillis, J. (2020). "Towards inferring cognitive state changes from pupil size variations in real world conditions," in *ACM Symposium on Eye Tracking Research and Applications* (virtual), 1–10. doi: 10.1145/3379155.3391319

Melamud, O., Goldberger, J., and Dagan, I. (2016). "context2vec: learning generic context embedding with bidirectional LSTM," in *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning* (Berlin), 51–61. doi: 10.18653/v1/K16-1006

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., and Hays, J. (2016). "Webgazer: Scalable webcam eye tracking using user interactions," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016* (New York City, NY). doi: 10.1145/2702613.2702627

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. 12, 2825–2830. doi: 10.5555/1953048.2078195

Peysakhovich, V., Causse, M., Scannella, S., and Dehais, F. (2015). Frequency analysis of a task-evoked pupillary response: luminance-independent

measure of mental effort. *Int. J. Psychophysiol.* 97, 30–37. doi: 10.1016/j.ijpsycho.2015.04.019

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., and Aguilar, M. (2010). Pupil dilation during visual target detection. *J. Vision* 10:3. doi: 10.1167/10.10.3

Raiturkar, P., Kleinsmith, A., Keil, A., Banerjee, A., and Jain, E. (2016). "Decoupling light reflex from pupillary dilation to measure emotional arousal in videos," in *Proceedings of the ACM Symposium on Applied Perception, SAP 2016* (Anaheim, CA), 89–96. doi: 10.1145/2931002.2931009

Sibley, C., Coyne, J., and Baldwin, C. (2011). "Pupil dilation as an index of learning," in *Proceedings of the Human Factors and Ergonomics Society* (Los Angeles, CA), 237–241. doi: 10.1177/1071181311551049

Slykhuis, D. A., Wiebe, E. N., and Annetta, L. A. (2005). Eye-tracking students' attention to powerpoint photographs in a science education setting. *J. Sci. Educ. Technol.* 14, 509–520. doi: 10.1007/s10956-005-0225-z

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM neural networks for language processing. *Interspeech* 2012, 194–197.

van den Brink, R. L., Murphy, P. R., and Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLoS ONE* 11:e0165274. doi: 10.1371/journal.pone.0165274

Wang, C.-A., Boehnke, S. E., White, B. J., and Munoz, D. P. (2012). Microstimulation of the monkey superior colliculus induces pupil dilation without evoking saccades. *J. Neurosci.* 32, 3629–3636. doi: 10.1523/JNEUROSCI.5512-11.2012

Wang, J. (2011). "Pupil dilation and eye tracking," in *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide*, eds M. Schulte-Mecklenbeck, A. Kuehberger, J. G. Johnson, and J. G. Johnson (New York, NY: Psychology Press New York), 185–204.

Wiedbusch, M. D., and Azevedo, R. (2020). "Modeling metacomprehension monitoring accuracy with eye gaze on informational content in a multimedia learning environment," in *ACM Symposium on Eye Tracking Research and Applications* (virtual), 1–9. doi: 10.1145/3379155.3391329

Winn, B., Whitaker, D., Elliott, D. B., and Phillips, N. J. (1994). Factors affecting light-adapted pupil size in normal human subjects. *Investig. Ophthalmol. Visual Sci.* 35, 1132–1137.

Xu, J., Wang, Y., Chen, F., and Choi, E. (2011). "Pupillary response based cognitive workload measurement under luminance changes," in *IFIP Conference on Human-Computer Interaction* (Lisbon), 178–185. doi: 10.1007/978-3-642-23771-3_14

Xu, S., Jiang, H., and Lau, F. C. (2008). "Personalized online document, image and video recommendation via commodity eye-tracking," in *RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems* (Lausanne), 83–90. doi: 10.1145/1454008.1454023

Conflict of Interest: SK and PA are founders at the company Fovea Inc. Fovea Inc. did not fund or take part in the experiment and analysis.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

METHODS
published: 17 March 2021
doi: 10.3389/fpsyg.2021.656388

Gaze Coordination of Groups in Dynamic Events – A Tool to Facilitate Analyses of Simultaneous Gazes Within a Team

*Frowin Fasold[1], André Nicklas[1], Florian Seifriz[1], Karsten Schul[1], Benjamin Noël[1], Paula Aschendorf[1] and Stefanie Klatt[1,2]**

[1] Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany, [2] Institute of Sports Science, University of Rostock, Rostock, Germany

The performance and the success of a group working as a team on a common goal depends on the individuals' skills and the collective coordination of their abilities. On a perceptual level, individual gaze behavior is reasonably well investigated. However, the coordination of visual skills in a team has been investigated only in laboratory studies and the practical examination and knowledge transfer to field studies or the applicability in real-life situations have so far been neglected. This is mainly due to the fact that a methodological approach along with a suitable evaluation procedure to analyze the gaze coordination within a team in highly dynamic events outside the lab, is still missing. Thus, this study was conducted to develop a tool to investigate the coordinated gaze behavior within a team of three human beings acting with a common goal in a dynamic real-world scenario. This team was a (three-person) basketball referee team adjudicating a game. Using mobile eye-tracking devices and an indigenously designed software tool for the simultaneous analysis of the gaze data of three participants, allowed, for the first time, the simultaneous investigation of the coordinated gaze behavior of three people in a highly dynamic setting. Overall, the study provides a new and innovative method to investigate the coordinated gaze behavior of a three-person team in specific tasks. This method is also applicable to investigate research questions about teams in dynamic real-world scenarios and get a deeper look at interactions and behavior patterns of human beings in group settings (for example, in team sports).

Keywords: team sports, officials, gaze behavior, teamwork, eye-tracking

OPEN ACCESS

Edited by:
Jonathan Touryan, United States Army Research Laboratory, United States

Reviewed by:
Qianwen Wan, Audi of America, United States
Mario Dalmaso, University of Padua, Italy

***Correspondence:**
Stefanie Klatt
s.klatt@dshs-koeln.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *20 January 2021*
Accepted: *24 February 2021*
Published: *17 March 2021*

Citation:
Fasold F, Nicklas A, Seifriz F, Schul K, Noël B, Aschendorf P and Klatt S (2021) Gaze Coordination of Groups in Dynamic Events – A Tool to Facilitate Analyses of Simultaneous Gazes Within a Team. Front. Psychol. 12:656388. doi: 10.3389/fpsyg.2021.656388

INTRODUCTION

If a group of human beings tries to fulfill the requirements of a specific task, e.g., to solve a problem, or to reach a specific performance criterion, the members of this group can coordinate their abilities and skills to optimize their behavior and/or performance. For instance, lifeguards coordinating scanning of water surfaces to minimize the risk of a failure to perceive a drowning person, or in a dangerous situation during rush hour traffic on the road, the role of co-driver in observing the traffic to minimize the risk of a road accident. Similarly, in team sports, coordination of visual skills

could be relevant for the overall performance of both, a team of sportspeople and/or referees.

The coordination of gaze among group members or within a team has almost exclusively been studied only in different experimental laboratory investigations. In a police-training simulation (on computer screens), for example, Neider et al. (2010) showed that different options for sharing information (verbal communication, shared gaze *via* gaze cursors of partners fixations), can influence performance in spatial referencing. The authors stated that verbal and visual information can be helpful for interpersonal coordination. An earlier study showed that in simple perceptual search tasks, the coordination of gazes through the observation of a partner's gaze leads to superior performances compared to individual search (Brennan et al., 2008). Bahrami et al. (2010) demonstrated in a low-level perceptual task (visual detection of a target stimulus) that team (dyads, i.e., two individuals) decisions are better than the decisions made by just one observer. A necessary condition for better performance is the opportunity of free communication within the dyads.

Other than the aforementioned studies, Macdonald and Tatler (2018) provided a description of the non-verbal communication cues that people use in real-world situations. They investigated the synchronized gaze behavior of a pair engaged in the activity of making a cake and observed how the roles of the subjects (chef and gatherer) influence the gaze behavior in this (social) interaction task (Macdonald and Tatler, 2018). Other than this study, almost no research concerning the interaction of gazes or the gaze coordination of people working in a team setting has been published so far. In particular, there has been no application-oriented method which is presented to analyze teams' gaze behavior, even though, the meta-analyses by Mann et al. (2007) as well as by Gegenfurtner et al. (2011) has suggested that specific cognitive abilities are better investigated in natural environments. Risko et al. (2016) additionally noted the absence of applicability of laboratory studies to real-world environments in the area of social attention.

Over the last few decades, technological advances in mobile eye tracking systems have made it easier to conduct field studies on gaze behavior [Kredel et al., 2017, also see Wan et al. (2019)]. In recent times, software solutions for (automated) data analyses have been widely available (e.g., Panetta et al., 2019, 2020). However, there is almost no procedure to evaluate the synchronized tracking of gaze behavior within teams (through eye tracking) in more or less dynamic situations. This shows that although the analysis of individual gaze behavior has been facilitated, it has not led to an increase of research on gaze behavior within groups.

Interestingly, this is also the case in research in areas like sports science and sports psychology, where eye tracking has become an often-used method to analyze athletes' or referees' gaze behavior. Recent reviews about gaze behavior in sports (Kredel et al., 2017; Hüttermann et al., 2018) have given an extensive overview on contents, methods, and practical applications which have been used in the past decades. Notably, all the studies included in the reviews, focused on individual parameters and joint activities of groups or team members. Considering that the performance in team sports (athletes and officials/referees) is primarily based on the coordination and interaction of individual skills performed by each team member, the lack of research on the subject is surprising.

However, there has been one study which investigated the coordinated gaze behavior of handball referees adjudicating a game (Fasold et al., 2018). While Fasold et al. (2018) showed that the analysis of the coordinated gaze of a dyad can be done well by using available software tools (e.g., eye-tracking device application, KINOVEA), it was evident that the observation of more than one person at the same time requires a well-structured experimental set-up. Furthermore, this procedure is associated with extensive processing of high volumes of data, especially considering that analyzing data of gaze behavior in highly dynamic situations is mainly done manually as a frame-by-frame analysis. Manual analysis is possibly the main reason for the research gap with regard to the analysis of the gaze behavior of two or more individuals in a natural environment. This is also related to the fact that a method to technically analyze the gazes of more than two people in a time-effective way is missing [considering that freeware solutions, most often, only allow the analysis of a dyad, see Fasold et al. (2018)].

The current study expands the approach of Fasold et al. (2018) by analyzing the synchronized gaze behavior of three individuals working together in a team. In contrast to lab-based studies, which allow high-frequency data collection and algorithmic based data analyses (Mele and Federici, 2012), in field studies it is often not possible to use automated analyses due to the dynamic movements, the accelerations, and the three-dimensional and ever changing areas of interest (AOI). While previous approaches for automated analyses of individual gaze behaviors in more or less dynamic environments do exist (e.g., Panetta et al., 2019), a manual frame-by-frame analysis is necessary to investigate the simultaneous gaze behavior of three referees. This kind of analysis, although time-consuming, is reliable and functional for sports practice (cf. Fasold et al., 2018). Therefore, the main focus of the current study was on the development of a reasonable way of analyzing data and figuring out a way to handle the large dataset (three eye-tracking devices).

Study

In investigating performances in team sports, usually, it is mainly the athletes who are the primary focus. But referees, who play a vital role in successfully conducting a game, have to deliver high performances in judging a large number of interactions in highly dynamic situations involving a lot of cues/athletes (e.g., Plessner and MacMahon, 2013). The relevance of the visual system in conducting refereeing tasks is well known, as documented by MacMahon et al. (2015, p. 47): "Perceiving, and in most cases seeing, is at the root of any judgment and decision in refereeing in sports." The assumption that groups of people can perceive more than an individual would, is derived from the refereeing regulations of various sports. The regulations of a multitude of (professional) sports demand the presence of a team of referees, rather than an individual referee (e.g., two referees in volleyball or team handball, three referees in basketball or ice hockey). Thus, for our study, we chose to

simultaneously observe the gaze behavior of three basketball referees by means of three individual mobile eye-tracking systems for each of them during a preparation game. In professional basketball, usually the 3-Person Officiating system (3PO) is used, meaning three referees are systematically assigned to the court in order to avoid as little action as possible (FIBA, 2016). Almost always, it's the FIBA which defines the zone/area which has to be covered by a specific referee (see **Figure 1**), and thus, it also determines how the referees should coordinate their gaze behavior (FIBA, 2016).

Despite these clear guidelines, it is actually unclear if a team of referees is able to simultaneously cover the court in the most optimum manner. To test the applicability of a new method to analyze the simultaneous gaze behavior of three people, we analyzed the referees' gaze behavior and compared it to the FIBA guidelines for refereeing. It is noteworthy that FIBA (2016) clarifies that the ground coverage by referees in an active game is not static and has to be adapted according to the dynamic flow of the game. Overlapping and/or dual coverage of some areas is not necessarily wrong and may be indicative of the functioning of the way that the referee team works.

METHODS

This study is the starting point of a larger project which evaluates the on-court visual behavior and social interaction among the referees in the team. The simultaneous coordination of gazes is one part of this application-oriented investigation, the other being the large amount of data resulting from such interaction. So far, a method to observe this interaction or measure the data has not been developed. Therefore, we developed a tool to facilitate a frame-by-frame analyses of simultaneous gaze behavior of several observers and tested it on a representative data set.

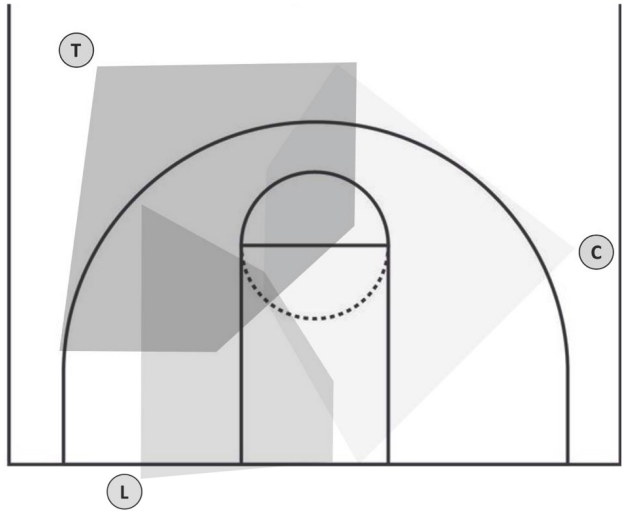


FIGURE 1 | Coverage of gaze areas of the three referees' positions (L, lead; T, trail; C, center) in the 3PO system, predefined by the guidelines for referees [modified figure according to FIBA (2016), p. 178)].

Tool Development

In order to analyze the simultaneous video streams (gaze data) of two people working as a team in a natural environment, a software named KINOVEA was applied in the study by Fasold et al. (2018). However, that method only allows the simultaneous analysis of two video streams. Moreover, while analyzing data, frame-by-frame video players' analysis is not sufficient because it only allows watching recordings of the several gazes frame-by-frame, without any options to save the results of such analyses. That is, all data must be separately entered and saved by means of an additional computer program. Such a procedure–switching from one video stream to another or watching several gaze recordings of the same situation in succession and switching from one computer program to another in order to enter data–is time-consuming and prone to errors.

Initially for the current project, but with a practical applicability beyond that for many future eye-tracking projects, we developed a tool enabling the simultaneous analysis of four video streams. This tool allows observation of three video streams of gaze data recorded with the eye-tracking devices (Referees A, B, and C) and a fourth stream showing the complete scene video, in the current case, recorded from the stand in the middle of the playing field (just for the validation of the game situations). This tool offers the possibility to play the videos frame by frame with the inserted gaze points extracted from the eye-tracking software (Pupil Player). We developed the tool using components of Delphi XE3, PasLibVlc[1] and the commercial TMS Component Studio[2]. The complete source code is available in a public GIT repository[3]. From the TMS Component Studio, we used only the components of the feature-rich user interface to create a modern user interface for our tool. Some features of the PasLibVlc are also used in the well-known VLC-Player. This ensures that for the purpose of our tool, most video formats can be played back seamlessly and this can also be used with other eye-tracking systems.

Our main goal for the software-user interface was to create a simple and efficient collection of the relevant data from various observers. Therefore, we developed a project structure to save all the necessary data of the used video files in one repository. The project properties included, among other things, the video file location, the position of each window of the video streams, and the synchronized start frame of the video streams. These properties were saved in a separate project file. After saving the files, an analysis could be started again after an interruption without much effort by simply loading the project property file. In order to avoid any data loss, the current status of the analysis was saved after editing each frame. For each gaze data stream, the analyst could choose the AOI of the referee who was being observed and see if the referee covered his/her primary AOI. Afterward, with a press of a button, save all entries could be saved into a csv file, followed by the next frame appearing automatically for all four videos (**Figure 2**). In the end, this csv file could be exported and used for further analyses.

[1]https://prog.olsztyn.pl/paslibvlc/

[2]https://www.tmssoftware.com

[3]https://github.com/Seifriz/AFVS

FIGURE 2 | Screenshot of the user interface of the analysis tool enabling the analysis of four video streams at the same time.

Case Study

Participants and Design

A single-case study with a team of expert referees was chosen for this study. This team (all males) adjudicated a preparation game of the German Basketball Bundesliga (Ref A, age = 29, experience on German expert level = 9; Ref B, age = 36, experience = 10; Ref C, age = 44, experience = 17; all with experience at the international level). All of them had normal or corrected-to-normal vision.

Ethics Statement

All participants provided written informed consent prior to their participation, in accordance to the principles outlined by the World Medical Association's Declaration of Helsinki and the Office of Research Ethics at the German Sport University Cologne (ethics proposal number: 141/2018). Participants were initially told that the study would investigate their communication strategy after a call during a game. After the game, they were informed about the real aim of the study.

Apparatus

Three Pupil Core mobile eye-tracking systems (Pupil Labs GmbH, Berlin, Germany) were used in this study. We used the binocular mobile eye-tracking headset connected to the mobile bundle which composed of a Motorola Moto Z2 or Z3 Play with an USBC-USBC cable. The eye movements were recorded with 200 Hz and were matched to the simultaneously captured scene videos recorded at 30 Hz (30 frames per second).

Procedure

Prior to testing, the referees were informed about the procedure. It was explained to them how eye tracking devices work, and they were given instructions on how to handle them. One hour before the game, the eye tracking headsets were adjusted to the participants. Thereafter, the referees conducted their normal pre-game warm-up without the eye tracking systems until 5 min before the start of the game. Then, the referees and examiners met in front of the scorer's table. The eye tracking systems were returned to the referees and the *Manual Marker Calibration* was conducted (Kassner et al., 2014). The referees then placed themselves on the free-throw line and looked toward the basket. One of the examiners held a Pupil Calibration Marker v0.4 in his/her hand and placed himself/herself 1 m in front of the referees. The referees were told not to move their heads and to follow the route of the marker only with their eyes. A predefined route was used to calibrate every single eye-tracker. At half-time, the recordings were stopped and 5 min before the second half, the same calibration procedure was repeated. Shortly before these calibration processes, the eye trackers were synchronized by time *via* a visual signal. The recordings were saved on an SD-card and calibrated *post hoc* using the Pupil Player application (version 1.17, Pupil Labs GmbH, Berlin, Germany; Kassner et al., 2014).

Data

To test the new developed analysis tool, this case study was conducted in line with a game specific task–in this case, *the behaviors of the referees in critical foul situations according the guidelines*. An expert–a 52-years-old basketball professional with experience as a player, a coach and a referee on national level–reviewed the whole game and rated 23 scenes as highly challenging foul situations with critical decisions. The scenes were shortlisted based on the choice of the expert on relevant time periods before the call/decision. For every frame, the experimenter determined whether the referees' gaze was inside the primary observation area (according to FIBA guidebook) or not. Furthermore, for every frame, the coordinated gaze behavior was assessed on the basis of whether the referees looked at the same AOI or not. The AOI were defined

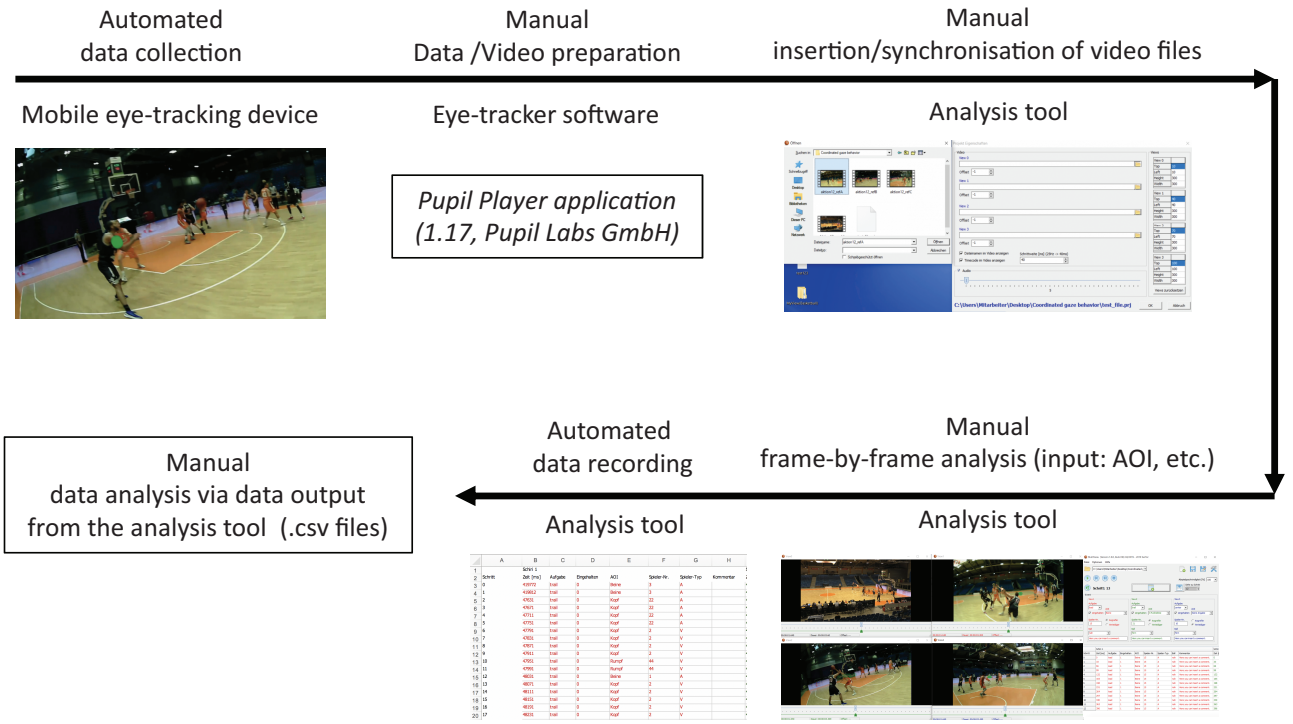


FIGURE 3 | Flow chart of the working process for analyzing gaze coordination in dynamic events using the new tool.

as the ball, the basket, another referee, the scorer's table, defending or attacking players, the zone, and the three-point line.

Data Analysis by Using the New Tool

Data Integration and Synchronization

After launching the analysis tool, the control (base) window and four other smaller windows (called view0, view1, view2, and view3) are displayed on the screen. After saving a new project, a window appears where the path name of the video files needs to be inserted. In addition to the selection of the video file, the analyst has to insert the videos' frame rate in ms (e.g., 40 ms = 25 Hz; 33.33 ms = 30 Hz). He/she can also select whether the audio track of the video file is played or not and the way the windows (view0–3) are arranged. The program also allows the synchronization of videos.

Steps of Analysis

After clicking on the *analysis button,* the program will start with the first frame. For the recordings of the gazes (of the referees), the appropriate AOI (ball, basket, scores table or shot clock) or player (attacking or defending, including the jersey number) can be chosen. In our study, we could also record if the referees covered the primary area of responsibility. Moreover, an additional insert field to add comments for each column is available. After the analyst completes entries for one frame, all the relevant information is saved automatically followed by access to the next frame. In the end, a csv file is available with all the data and can be exported and used for further analyses. **Figure 3** represents a flowchart of the working process, highlighting how the simultaneous analysis of multiple persons' gaze behavior is possible using the new tool in dynamic events.

RESULTS AND DISCUSSION

The primary aim of this study was to develop and apply, for the first time, a method to investigate the simultaneous gaze behavior of a team of three people operating in a highly dynamic environment. One major challenge in such an investigation is the large amount of data generated. Our method provides a solution for this problem by presenting a procedure which allows easy classification of large datasets and subsequently, makes analyses easier.

The method we have developed is not only economical, but also offers a potential for wide applicability. Using mobile eye-tracking devices in highly dynamic and specialized situations could affect the natural behavior of the participants. However, in our study, the participants did not report any sizeable constraints during the process of data collection despite the game activity. They all reported that they needed a short period of time to familiarize themselves with the eye-tracking device and after a few minutes, they didn't feel any constraints.

The recording of the gaze behaviors of the participants during the chosen scenes within the basketball game, did not

present any major problems either and the gaze points of all the participants were clearly visible. The use of the analysis tool makes frame-by-frame analysis of gaze recordings of multiple participants feasible, because all the three videos (of the referees) could be analyzed simultaneously, and no additional computer program was needed for manually transferring the data from the recorded videos. Previous studies have presented approaches to analyze simultaneous gazes of dyads in laboratory studies (e.g., Neider et al., 2010) or the simultaneous gazes of even bigger groups in a digital classroom (Nyström et al., 2017). The highly technical demands of investigating gaze of a group has been the focus of these investigations and past results have shown that such paradigms work well with acceptable deviations in different technical parameters (e.g., latency, Nyström et al., 2017). Based on these results, Scurr et al. (2014) have stated that several studies show methodologies for simultaneous gaze analysis, but none of them has been applied in highly dynamic situations. As an innovative extension of these investigations and research methods, the methodological approach presented in this study, allows an analysis of teams' gaze behavior in dynamic actions. Even beyond that, it can be applied to other research fields, such as social cognition in which the presence of another person has been found to affect the gaze of an observer (e.g., Nasiopoulos et al., 2015); or this approach could help to assess how social attention is distributed in multi-agent contexts (cf. Dalmaso et al., 2020).

Results of the case study: The analysis included 2,339 video frames synchronized per referee. The mean duration of the scenes lasted 106.31 frames (*SD* = 40.36). In 47.07% of the analyzed scenes, the referees looked at a point in their primary observation area (with a high *SD* = 24.35%). This distribution of the coverage of the primary observation area varied as a function of the referees' positions (trail 54.58%, *SD* = 21.30; lead 39.03%, *SD* = 24.85; center 47.67, *SD* = 26.14). Analyzing the coordination of the referees' gaze, in just 5.61% of all analyzed frames, all three referees fixated their gaze on the same area. In 31.94% of all analyzed frames, two referees fixated their gaze on the same area (center + lead 13.34%; center + trail 10.90%; lead + trail 7.70%). The results showed that the referee team under observation followed the FIBA guidelines in approximately half of all included cases (playing situations). Interestingly, in majority of the analyzed frames (>90%), they distributed their gaze to different AOI to cover as many aspects of game actions as possible. In contrast, the single case study of Fasold et al. (2018) showed that the novice referees in handball did not coordinate their gaze behavior in an appropriate manner, i.e., both referees focused on areas close to the ball in about 80% of the analyzed data. Considering that in our study the tested team of referees was highly experienced, it does seem surprising that deviations from the guidelines occur here as well.

CONCLUSION

The method of analyzing synchronized gaze behavior of three individuals seems promising for future efforts in various other scenarios involving many individuals who try to work together as a team (e.g., personal protection, traffic monitoring). Furthermore, it could be used to replicate findings of basic experimental research on gaze behavior of teams (e.g., Bahrami et al., 2010; Neider et al., 2010) in natural environments or settings.

We were able to show that similar simultaneous gaze evaluation with more than one or two participants is possible in dynamic situations, enabling new possibilities in studying social and functional coordinated interactions in future. We extended the approach by Fasold et al. (2018) to three referees working together as a team, with the aim of perceiving as much of the relevant game actions as possible in a very information-rich environment. This study shows that using mobile eye-tracking devices (Pupil Labs) and a new analysis tool does make simultaneous gaze analyses in dynamic environments possible and even more efficient. To develop the analysis tool we utilized commonly used components (e.g., PubLibVlc) because these components guarantee high performance, stability and compatibility with all available video formats. The manual frame-by-frame analysis is still time consuming but, it can be conducted much faster considering the availability of software that facilitates the analytical process. This manual method may be necessary until machine-learning processes allow algorithmic-based data analysis in highly dynamic scenarios. In our case, we expect that the manual analysis will allow the report of frequencies next as a qualitative observation of the recorded data (game actions) and this could result in a step toward automated analysis. Although, the tool we developed is available on GIThub, we recommend adjusting the program to the specific needs of every researcher's own project. Furthermore, we suggest advancements to the new analysis tool by integrating keyboard short cuts or the development of a specialized keyboard like the ones used for video editing. These adaptations could further optimize the user interface of the analysis tool and could optimize manual data analysis.

To conclude, the current status of our software should be seen as a starting point for investigations into coordination of visual skills in groups, not only in sports, but also in everyday tasks where acting as a team is required to achieve the defined objectives.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FF, KS, BN, and SK developed the study concept and contributed to the design. AN collected the data and was supported by FF, KS, BN, and SK. AN analyzed the data together with FF, FS, KS, BN, and SK. FS developed the analysis tool. FF and SK wrote

the first draft of the manuscript. All authors helped to edit and revise the manuscript and approved the final submitted version of the manuscript.

FUNDING

This project was supported by a grant from the easyCredit Basketball Bundesliga. The authors also acknowledge financial support by the German Research Foundation and the University of Rostock within the funding program Open Access Publishing.

ACKNOWLEDGMENTS

The authors would like to thank Ricardo Wilhelm and Alessa Schwarting for collecting parts of the data.

REFERENCES

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. *Science* 329, 1081–1085. doi: 10.1126/science.1185718

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106, 1465–1477. doi: 10.1016/j.cognition.2007.05.012

Dalmaso, M., Castelli, L., and Galfano, G. (2020). Social modulators of gaze-mediated orienting of attention: a review. *Psychon. Bull. Rev.* 27, 833–855. doi: 10.3758/s13423-020-01730-x

Fasold, F., Noël, B., Wolf, F., and Hüttermann, S. (2018). Coordinated gaze behaviour of handball referees: a practical exploration with focus on the methodical implementation. *Mov. Sport Sci. Sci. Motricité* 102, 71–79. doi: 10.1051/sm/2018029

FIBA (2016). *FRIP Level 1. Home study bookl*. Mies: FIBA. Available online at: http://www.basketref.com/documents/fiba_materials_2018/New_2017_FRIP%20L1_Home_Study_Book_4_Refereeing_v1.0.pdf

Gegenfurtner, A., Lehtinen, E., and Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educ. Psychol. Rev.* 23, 523–552. doi: 10.1007/s10648-011-9174-7

Hüttermann, S., Noël, B., and Memmert, D. (2018). Eye tracking in high-performance sports: evaluation of its application in expert athletes. *Int. J. Comput. Sci. Sport* 17, 182–203. doi: 10.2478/ijcss-2018-0011

Kassner, M., Patera, W., and Bulling, A. (2014). "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, (New York, NY: ACM), 1151–1160.

Kredel, R., Vater, C., Klostermann, A., and Hossner, E.-J. (2017). Eye-Tracking technology and the dynamics of natural gaze behavior in sports: a systematic review of 40 years of research. *Front. Psychol.* 8:1845. doi: 10.3389/fpsyg.2017.01845

Macdonald, R. G., and Tatler, B. W. (2018). Gaze in a real-world social interaction: a dual eye-tracking study. *Q. J. Exp. Psychol.* 71, 2162–2173. doi: 10.1177/1747021817739221

MacMahon, C., Mascarenha, D., Plessner, H., Pizzerra, A., Oudejans, R. R. D., and Raab, M. (2015). *Sports Officials and Officiating: Science and Practice*. Abingon: Routledge.

Mann, D. T. Y., Williams, A. M., Ward, P., and Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: a meta-analysis. *J. Sport Exerc. Psychol.* 29, 457–478. doi: 10.1123/jsep.29.4.457

Mele, M. L., and Federici, S. (2012). Gaze and eye-tracking solutions for psychological research. *Cogn. Process.* 13, 261–265. doi: 10.1007/s10339-012-0499-z

Nasiopoulos, E., Risko, E. F., and Kingstone, A. (2015). "Social attention, social presence, and the dual function of gaze," in *The Many Faces of Social Attention*, eds A. Puce and B. Bertenthal (Cham: Springer), 129–155.

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., and Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychon. Bull. Rev.* 17, 718–724. doi: 10.3758/PBR.17.5.718

Nyström, M., Niehorster, D. C., Cornelissen, T., and Garde, H. (2017). Real-time sharing of gaze data between multiple eye trackers–evaluation, tools, and advice. *Behav. Res. Methods* 49, 1310–1322. doi: 10.3758/s13428-016-0806-1

Panetta, K., Wan, Q., Kaszowska, A., Taylor, H. A., and Agaian, S. (2019). Software architecture for automating cognitive science eye-tracking data analysis and object annotation. *IEEE Trans. Hum. Mach. Syst.* 49, 268–277.

Panetta, K., Wan, Q., Rajeev, S., Kaszowska, A., Gardony, A. L., Naranjo, K., et al. (2020). Iseecolor: method for advanced visual analytics of eye tracking data. *IEEE Access* 8, 52278–52287. doi: 10.1109/ACCESS.2020.2980901

Plessner, H., and MacMahon, C. (2013). "The sport official in research and practice," in *Developing Sport Expertise: Researchers and Coaches put Theory into Practice*, 2nd Edn, eds D. Farrow, J. Baker, and C. MacMahon (London: Routledge), 71–92.

Risko, E. F., Richardson, D. C., and Kingstone, A. (2016). Breaking the fourth wall of cognitive science: real-world social attention and the dual function of gaze. *Curr. Dir. Psychol. Sci.* 25, 70–74. doi: 10.1177/0963721415617806

Scurr, J. C., Page, J., and Lunt, H. (2014). A methodological framework for capturing relative eyetracking coordinate data to determine gaze patterns and fixations from two or more observers. *Beh. Res. Methods* 46, 922–934. doi: 10.3758/s13428-014-0443-5

Wan, Q., Kaszowska, A., Panetta, K., Taylor, H. A., and Agaian, S. (2019). A comprehensive head-mounted eye tracking review: software solutions, applications, and challenges. *Electron. Imaging* 2019, 654–651. doi: 10.2352/ISSN.2470-1173.2019.3.SDA-654

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 24 May 2021
doi: 10.3389/fnins.2021.656913

Gaze During Locomotion in Virtual Reality and the Real World

Jan Drewes[1,2], Sascha Feder[3] and Wolfgang Einhäuser[2]**

[1] Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu, China, [2] Physics of Cognition Group, Institute of Physics, Chemnitz University of Technology, Chemnitz, Germany, [3] Cognitive Systems Lab, Institute of Physics, Chemnitz University of Technology, Chemnitz, Germany

How vision guides gaze in realistic settings has been researched for decades. Human gaze behavior is typically measured in laboratory settings that are well controlled but feature-reduced and movement-constrained, in sharp contrast to real-life gaze control that combines eye, head, and body movements. Previous real-world research has shown environmental factors such as terrain difficulty to affect gaze; however, real-world settings are difficult to control or replicate. Virtual reality (VR) offers the experimental control of a laboratory, yet approximates freedom and visual complexity of the real world (RW). We measured gaze data in 8 healthy young adults during walking in the RW and simulated locomotion in VR. Participants walked along a pre-defined path inside an office building, which included different terrains such as long corridors and flights of stairs. In VR, participants followed the same path in a detailed virtual reconstruction of the building. We devised a novel hybrid control strategy for movement in VR: participants did not actually translate: forward movements were controlled by a hand-held device, rotational movements were executed physically and transferred to the VR. We found significant effects of terrain type (flat corridor, staircase up, and staircase down) on gaze direction, on the spatial spread of gaze direction, and on the angular distribution of gaze-direction changes. The factor world (RW and VR) affected the angular distribution of gaze-direction changes, saccade frequency, and head-centered vertical gaze direction. The latter effect vanished when referencing gaze to a world-fixed coordinate system, and was likely due to specifics of headset placement, which cannot confound any other analyzed measure. Importantly, we did not observe a significant interaction between the factors world and terrain for any of the tested measures. This indicates that differences between terrain types are not modulated by the world. The overall dwell time on navigational markers did not differ between worlds. The similar dependence of gaze behavior on terrain in the RW and in VR indicates that our VR captures real-world constraints remarkably well. High-fidelity VR combined with naturalistic movement control therefore has the potential to narrow the gap between the experimental control of a lab and ecologically valid settings.

Keywords: gaze, eye tracking, virtual reality, real world, virtual locomotion

OPEN ACCESS

Edited by:
Barry Giesbrecht, University of California, Santa Barbara, United States

Reviewed by:
Jonathan Touryan, United States Army Research Laboratory, United States
Emilia Biffi, Eugenio Medea (IRCCS), Italy

****Correspondence:***
Jan Drewes
mail@jandrewes.de
Wolfgang Einhäuser
wolfgang.einhaeuser-treyer@physik.tu-chemnitz.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *21 January 2021*
Accepted: *27 April 2021*
Published: *24 May 2021*

Citation:
Drewes J, Feder S and Einhäuser W (2021) Gaze During Locomotion in Virtual Reality and the Real World. Front. Neurosci. 15:656913. doi: 10.3389/fnins.2021.656913

INTRODUCTION

The question what guides our gaze in realistic settings has been of interest to researchers for decades. Since the pioneering studies of Buswell (1935) and Yarbus (1967), this issue has long been reduced to eye movements during scene viewing; that is, observers looking at pictures of natural scenes with little to no head and body movements. Since the turn of the millennium, many computational models that predict gaze allocation for such scene viewing have been developed. Since, Itti et al. (1998) adapted Koch and Ullman (1985) "saliency map" to predict fixated locations in a natural scene, many models followed the idea to combine (low-level) image features using increasingly sophisticated schemes or optimality principles (e.g., Bruce and Tsotsos, 2006; Harel et al., 2006; Zhang et al., 2008; Garcia-Diaz et al., 2012). As such models presumably built-in some implicit (proto-)object representation and objects are crucial for gaze guidance (Stoll et al., 2015), it comes as no surprise that models that use deep neural networks that share their lower-levels with object recognition models (e.g., Kümmerer et al., 2015), have become most successful and close to the theoretical image-computable optimum in predicting gaze during free viewing of natural scenes. However, such image-computable models do not explicitly include other factors that are crucial for gaze guidance in natural scenes (Tatler et al., 2011), such as the task (Buswell, 1935; Yarbus, 1967; Hayhoe and Ballard, 2005; Underwood and Foulsham, 2006; Henderson et al., 2007; Einhäuser et al., 2008), semantics (Henderson and Hayes, 2017) or interindividual differences (de Haas et al., 2019). Crucially, most modeling and experimental studies alike, have used scene viewing with the head-fixed, which provides good experimental control, but the transfer to real-world scenarios is less clear.

In typical laboratory settings, where the movement of head and body is highly constrained, eye movements typically consist mainly of saccades – rapid shifts of gaze – and fixations – times in which the eyes are relatively stable and only small fixational eye movements [drift, microsaccades and tremor, Rolfs (2009) and Martinez-Conde et al. (2004) for reviews] occur. When a target moves though the visual field, it can be followed by smooth pursuit eye movements (Ilg, 2002; Spering and Montagnini, 2011); when the whole visual field moves, an optokinetic nystagmus is induced that stabilizes the image on the retina through slow eye-movement phases, whose dynamics is similar to pursuit (Magnusson et al., 1986), and resets the eyes in their orbit by fast phases, whose dynamics is similar to saccades (e.g., Garbutt et al., 2001). If the head is moved, the vestibular ocular reflex (VOR) quickly stabilizes gaze by counterrotating the eyes relative to the head (Fetter, 2007). While these classes of eye movements can be distinguished based on their dynamics and use in part different neuronal circuitry (Ilg, 1997; Kowler, 2011 for reviews), during real-world behaviors these movements interact and their conceptual separation becomes less clear (Steinman and Collewijn, 1980). For example, if an observer tracks an object that is stationary in the world while they are moving in the world, conceptually, this would be close to a fixation, while the eyes are clearly moving relative to their orbit. Hence for complex scenarios it is often critical to carefully distinguish between separate coordinate systems (e.g., eye movements relative to the head – hereafter referred to as "eye-in-head," head movements relative to the world – "head in world," or eye movements relative to the world, hereafter "gaze-in-world") and to consider variables of interest, such as eye movement velocity in either coordinate frame, rather as a continuum than as means of distinguishing eye-movement classes strictly. Nonetheless, we still identify saccades based on velocity criteria (Engbert and Kliegl, 2003) for analysis purposes, while we do not separate any other classes further. Besides the mentioned convergent eye movements (both eyes move in unison), there are also divergent (vergence) eye movements, which we do not consider here, as in most cases objects of interest are at a considerable distance making the size of vergence movements small to negligible relative to other movements.

Even without an explicit task, participants exploring the real world (RW) at least need to navigate their environment and maintain a stable gait. Indeed, eye-movement behavior differs qualitatively, when walking through a natural world as compared to watching the same visual input as videos or series of stills with the head fixed ('t Hart et al., 2009; Foulsham et al., 2011). Moreover, gaze is affected by the difficulty of the terrain to be negotiated ('t Hart and Einhäuser, 2012; Thomas et al., 2020) and critical to guide an individual's steps (Matthis et al., 2018). Consequently, the constraints and implicit tasks imposed by the environment along with the freedom to move not only the eyes but also the head and the body to allocate gaze, limit the transfer from laboratory studies to real-world settings. At the same time, when aiming for general results beyond a specific application setting – such as sports (e.g., Land and McLeod, 2000; Hayhoe et al., 2012, for a review see Kredel et al., 2017), interface design (Thoma and Dodd, 2019), customer evaluation (Zhang et al., 2018) or driving (Land, 1992; Chapman and Underwood, 1998; Kapitaniak et al., 2015) to name just a few areas where eye-tracking has become a widely used tool – the degree of experimental control in a real-world setting is severely limited. This may become even more crucial when specific tasks such as search shall be studied, rather than free exploration or free viewing. Here, virtual reality (VR) has recently emerged as a viable alternative to overcome the gap between the limited ecological validity of the lab and the limited experimental control of the "wild."

VR, especially when displayed through head-mounted displays (HMDs) has some intrinsic limitations, such as a restricted field of view or limited resolution. Moreover, physiological factors such as the vergence/accommodation conflict (Kramida, 2016; Iskander et al., 2019), may lead to increased visual stress (Mon-Williams et al., 1998). However, thanks to ever improving display technology, decreasing costs and ease-of-use, over the recent years, VR systems have become a research tool in many fields. This includes – besides the entertainment market – highly regulated fields like medicine [e.g., Dentistry (Huang et al., 2018), education and training (Bernardo, 2017; Izard et al., 2018), simulation, diagnosis and rehabilitation of visual impairments (Baheux et al., 2005;

Jones et al., 2020)] and psychotherapy (e.g., autism therapy: Georgescu et al., 2014; Lahiri et al., 2015, fear and anxiety disorders: Hong et al., 2017; Kim et al., 2018; Matthis et al., 2018), as well as in areas directly relevant to psychophysical research such as attentional allocation (Helbing et al., 2020). As fears of long-term negative effects of VR use have so far not been confirmed (e.g., Turnbull and Phillips, 2017), and the recent VR goggles approach photorealistic capabilities while being more and more comfortable to wear, we are now in a position to ask, to what extent a HMD can be used as a proxy for a real-world setting in the context of gaze tracking – a question that has previously only been addressed in a limited scope. Pioneering the use of VR in eye-tracking research, Rothkopf et al. (2007) and Rothkopf and Ballard (2009) demonstrated that with identical visual environments the task – in their case collecting or avoiding obstacles – drastically alters gaze behavior relative to the objects of relevance. Meißner et al. (2017) made use of VR-based gaze tracking in the context of an augmented-reality shopping experience. Anderson et al. (2020) showed that both hand movements and gaze behavior in VR follow the same principles as in real life, at least while watching static natural scenes. VR and gaze tracking are also seeing widespread use in the field of driving simulation, allowing for test scenarios that would be dangerous or difficult to realize in the RW (e.g., Konstantopoulos et al., 2010; Zhang et al., 2017; Swan et al., 2020).

In spite of the increasing use of VR as display technology for eye-tracking experiments, the question as to how faithfully a VR setting mimics real-world constraints with respect to gaze allocation has remained largely unaddressed. Here, we address this issue for walking through a virtual and a real space. For such a direct comparison between gaze allocation when walking through the real and the virtual world, however, participants need to be tested in a sufficiently complex and large environment to allow actual locomotion, which needs to be closely and faithfully matched by a virtual copy of the same environment.

In the present study, we compare gaze while walking on a pre-defined path through three storeys of an office building to gaze while moving on a virtual high-fidelity copy of the same path (**Figure 1** and sample videos in the **Supplementary Material**). In VR, participants control their translational movement by a handheld controller (and do not actually translate), while they do execute rotational movements that are transformed into the matching rotational movement in the VR. We predefine different zones on the path (factor "sector" with levels "corridors," "ascending stairs," and "descending stairs") and assess robust measures of eye-movement behavior for both the RW and the VR (factor "world" with levels "VR" and "RW"). Assuming that the movement in VR is a good proxy for locomotion in the RW (with respect to gaze measurements), we would expect that differences in these measures found in the RW are also found in the VR, and remain largely unaffected by the choice of world. That is, under the hypothesis that VR faithfully approximates the RW, we expect main effects of the factor sector, but no interaction between sector and world for dependent variables characterizing relevant aspects of gaze allocation.

MATERIALS AND METHODS

Comparing Different Worlds

To compare gaze behavior between VR and RW, it is desirable to expose the participants to VR-generated surroundings that are as closely matched to the RW surroundings as possible. For practical reasons, the Physics building of Chemnitz University of Technology was chosen as the real-world location for this study and also modeled in VR.

Real World

Participants were instructed to walk through the building on a pre-defined route (**Figure 1**) at "their usual walking speed without unnecessary stopping." To enable participants to follow the route without actively engaging them at every turn, landmarks were placed at critical spots pointing in the correct direction. To avoid making the landmarks overly salient by falling out of the building context, a type of office chair abundantly available throughout the building was chosen. A plain white A4-sized paper with a black printed arrow was attached to the backrest, pointing the way (**Figure 2**).

The route started in the basement in the laboratory's commons area, and went through several corridors, lobbies and staircases until it returned on a different route to the same commons area. A detailed route description can be seen from **Figure 1**, including the segmentation into types of sectors (terrain) for the analysis. The route was inspected before each session, unforeseen obstacles were removed and any doors possibly interfering with the route were blocked open, such that participants did not have to interact with any object in their path. At the end of each session, the route was inspected again, and each recorded scene cam video was also manually inspected for such anomalies.

The experiment was conducted in the early evening hours of the European summer (ca. 18–21 h CEST), as these hours afforded both good natural illumination and minimal traffic within the building. Nonetheless, on occasion there were unforeseen obstacles in the path of the participant, and in 10 instances (max. 1 per individual) participants encountered other persons or doors not part of the walking route left open during the trial. Even though participants reported to believe that those incidents were part of the experiment, the corresponding periods during which the disturbance persisted (i. e., was visible) were excluded from gaze-data analysis to avoid data contamination.

Virtual Reality

To achieve best possible comparability between VR and reality, a high-fidelity 3D model of the entire physics building was developed, allowing identical walking routes in VR as well as reality (for a sample screenshot comparing RW and VR, see **Figure 2**; sample movies depicting both virtual and RW are available as **Supplementary Material**[1]). No human-like characters or avatars were placed in the virtual scenario. The software packages Blender (v. 2.79) and Unity (v. 2018.3.0f2)

[1]High quality versions of the videos in the Supplementary Material can be found at https://doi.org/10.6084/m9.figshare.14553738 and https://doi.org/10.6084/m9.figshare.14553759.

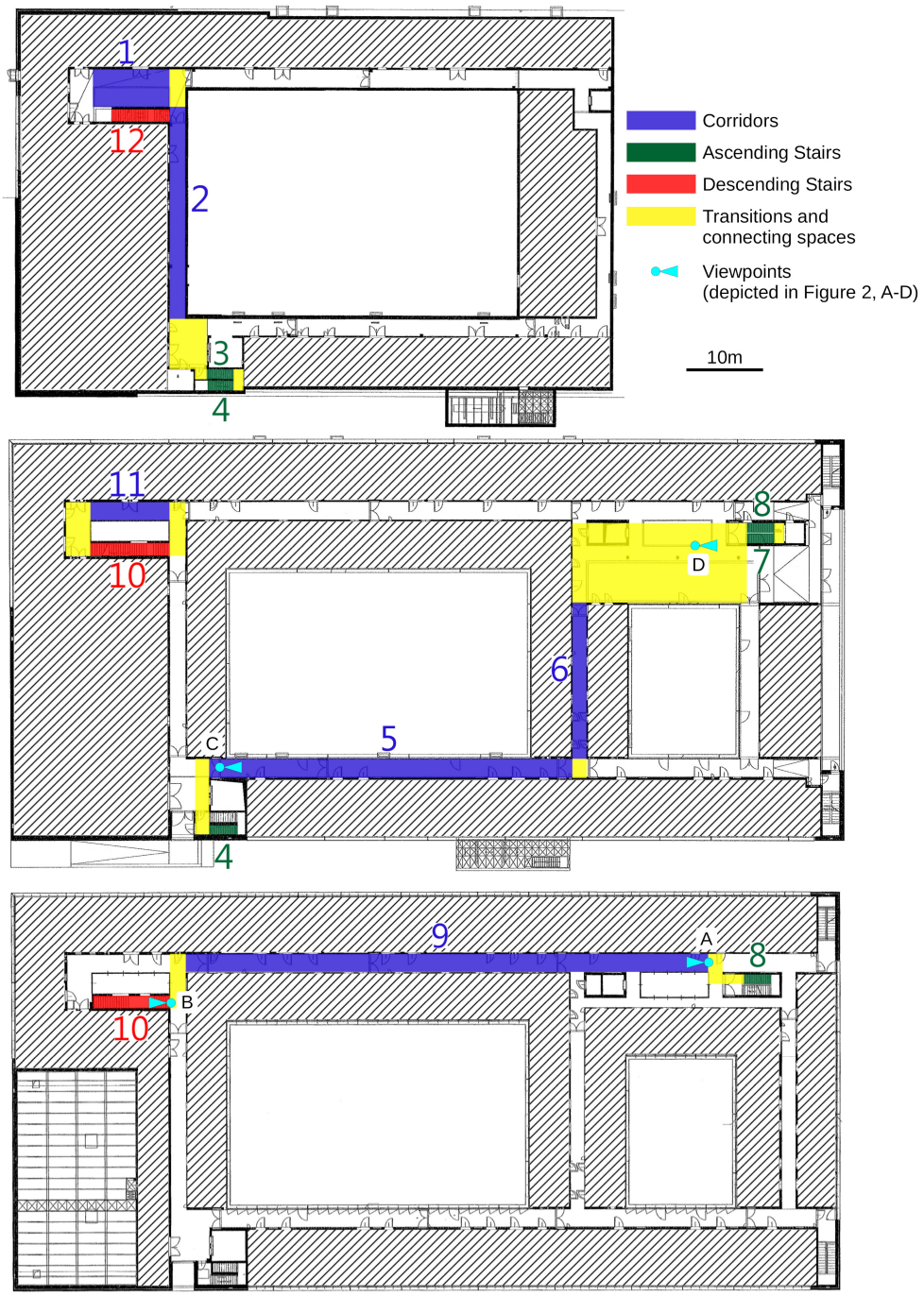


FIGURE 1 | Layout of the three floors (**top panel:** basement, **middle panel:** street level; and **bottom panel:** one floor above street level); of the experimental site and the selected walkway. Different sections are classified by color coding (see legend) and numbered in order of passing. Interior walls and structures irrelevant to the experiment grayed out for data-protection reasons. Map is to scale, scale bar corresponds to 10 m.

were used for developing and rendering the VR model. The internal details of the building were represented in the VR with great attention to detail and quality, including physical objects such as door handles, fire extinguishers, air vents, plants, and readable posters and showcases with objects inside. As a result, the virtual environment consists of 1,607 objects, whose total polygon number adds up to 4,453,370. Three hundred and forty different materials were used to texturize these objects. To limit the hardware load caused by the high polygon count of the model, a combination of culling operations offered by the Unity engine were used to minimize drawing operations without compromising visual quality. The main components of

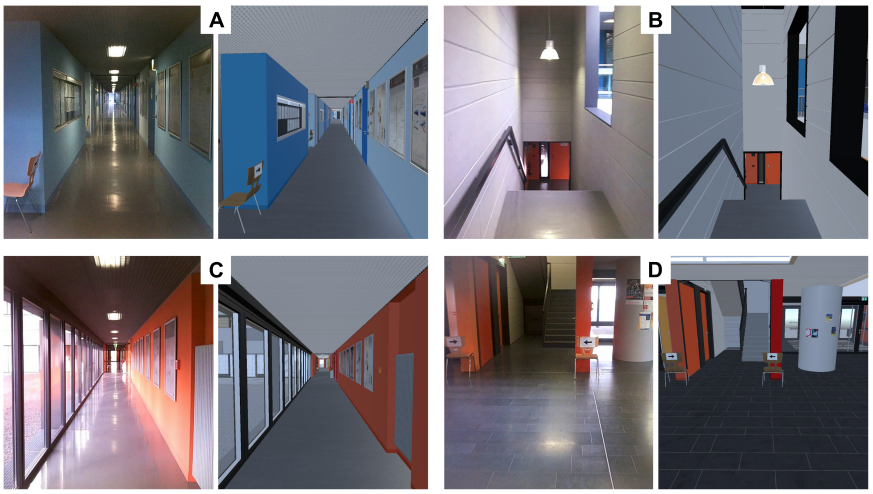


FIGURE 2 | Comparison of simulation and reality. Pairs of corresponding sample frames taken from the real world **(left)** and the virtual reality **(right)** recordings. VR rendering quality settings were as used during the experiment, (image quality can appear lower in the article pdf due to compression) RW were cut from the SMI glasses scene-camera recordings. **(A)** A corridor without windows. **(B)** Descending staircase. **(C)** A corridor with windows. **(D)** Parts of the lobby (not used in the analysis) and the entrance to an ascending staircase.

the model are the publicly accessible areas (corridors, staircases) of the reference building. They extend over five floors, which are connected by 242 steps and a virtual elevator that stops on four floors (elevator was not used in the present experiment). There are more than 200 doors and a similar number of scientific posters located along the corridors. Besides the public areas of the building, the laboratory where the virtual part of the experiment took place was also modeled. It served as starting and ending point of the predefined path in the experiment and made it easier for the test persons to switch between the RW and the virtual environment, because they started (finished) in the same position in VR where they put on (off) the headset in the RW. One seminar room and one office also were modeled in their entirety – for demonstration purposes and for use in follow-up experiments.

Matching Simulation and Reality

Within the VR, the height of origin of the participant's field of view (i.e., the virtual camera position) was adjusted individually to the physical height of each participant, to optimally match the visual appearances of the virtual and the RW. Proper camera height is also helpful to assist participants in fully immersing themselves in the VR, including the perceived ownership of their virtual bodies (van der Veer et al., 2019). The route that participants were to take during the experiment was marked with virtual copies of the marker chairs described above. Every chair's location and orientation from the real-world trials were copied faithfully to the VR, including the attached paper with the printed arrows (**Figure 2**).

Navigation in VR

Participants viewed the virtual building from a first-person perspective while moving their virtual body ("avatar") through the building. To avoid cyber sickness (or motion sickness, Golding and Gresty, 2005; Mazloumi Gavgani et al., 2018) while maintaining a naturalistic mode of navigation, a hybrid between physical tracking and joypad navigation was implemented. Forward (and if needed, backward) movement was controlled by means of the joypad on top of a VR controller. Rotational movements, in contrast, were actually executed by the participant and transformed to VR by tracking the VR controller held close to the body. When a participant pressed forward, the avatar would accelerate smoothly to a top speed of 1 m/s (equivalent) in the direction the participant's body was facing. Participants were instructed to rotate their whole body (and thus the VR controller with it) to determine the orientation of their avatar in the VR world. Through this, head movements were independent of motion direction, allowing for natural viewing behavior while at the same time allowing for an intuitive, semi-naturalistic navigation through the VR space. The top speed of 1 m/s was chosen to minimize the probability of motion/simulator sickness during the course of the experiment. The overall walking time was also relatively short (expected < 10 min), which should also help to avoid cyber sickness during the course of the experiment (Dużmańska et al., 2018; Mazloumi Gavgani et al., 2018). Indeed, no cases of cyber sickness were reported by the participants.

Experimental Setup and Gaze Recording

Real World

Real-world eye tracking was performed with a wearable eye tracker manufactured by SensoMotoric Instruments (SMI Eyetracking Glasses, ETG 2.1). Gaze data and scene camera video were recorded with a specially modified cell phone (Samsung Galaxy S5), which participants carried in a belt pocket. Gaze data were initially recorded at 60 Hz, while scene video was recorded at 25 Hz at a manufacturer-defined field of view of 60° horizontally and 46° vertically, which corresponds to the

gaze tracking range of the device. The manufacturer's built-in calibration was used to achieve a correct mapping between gaze and scene video. Calibration markers (3 × 3, spaced at 10° horizontally and vertically) were attached to a wall, and participants were instructed to fixate each marker for at least 2 s, once before walking along the path for manual inspection of the calibration prior to recording as well as once at the end of the experiment to allow for drift estimation.

Virtual Reality

For VR presentation and interaction, an HTC Vive VR headset was used in combination with a Vive hand controller. The headset offers a stereo display with a physical resolution of 1,080 × 1,200 pixels per eye at a refresh rate of 90 Hz, and a field of view of approx. 100° horizontally and 110° vertically. Position and rotation of the headset in space is tracked by means of 2 "Lighthouses" (laser scanners). While the VR and tracking capabilities of the headset were unchanged from the standard commercial package, eye tracking was realized through an aftermarket add-on manufactured by Pupil-Labs (Pupil Labs GmbH, Berlin, using Pupil Capture v. 1.11-4), consisting of two infrared cameras mounted inside the head set, tracking one eye each with a nominal frequency of 120 Hz at a camera resolution of 640 × 480 pixels. Eye tracking and VR computations were performed on a laptop (ASUS GL502VS, Intel Core i7 6700HQ, Nvidia GeForce GTX1070 GPU), allowing for time-synchronized data recording of both VR and eye/gaze events. The cables leading to the VR headset were loosely suspended from the ceiling above the participant to avoid exerting forces on the participants' head and neck.

To achieve a correct mapping between measured pupil position and gaze position in VR, the built-in calibration routine of the Pupil Labs eye tracker was followed by a custom calibration sequence, consisting of 3 × 3 calibration markers positioned at a grid spacing of 10° vertically and 11.5° horizontally. Participants were cued which marker to fixate by a change in marker color, and were requested to maintain fixation until the next marker was highlighted in random order by the operator (at least 3 s of fixation time each). The recorded raw data was then projected onto the known positions of the calibration grid using a 2-dimensional polynomial fitting procedure (Drewes et al., 2014). At the end of the VR recording session, the procedure was repeated to allow for drift evaluation.

Procedure and Participants

The order of conditions (VR vs. RW) was balanced across participants. Including briefing, data collection and debriefing, the experiment lasted about 40 min, depending on individual walking speed. Twelve individuals participated in the experiment (9 women, 2 men, and 1 unreported) with an average age of 22.8 years (18–33). Visual acuity was tested by means of a Snellen chart; all participants reported to be healthy and being able to walk and climb stairs without any restrictions or aid. Participants were explicitly instructed prior to the experiment that they should abort the experiment, if they experienced any motion sickness or discomfort; when asked informally in debriefing, no one indicated any signs of either motion sickness or discomfort. Participants were remunerated for their participation by 6€/h or course credit.

All procedures were performed in accordance with the Declaration of Helsinki, and were evaluated by the applicable ethics board (*Ethikkommission der Fakultät HSW, TU Chemnitz*) who ruled that no in-depth evaluation was necessary (case-no.: V-274-PHKP-WET-Augenb-11062018).

Data Processing and Analysis

For eye movements recorded in VR, as a first step the calibration solution as described above was applied. Three participants had to be excluded, as data quality did not allow for proper calibration. For one additional participant, data recording failed due to technical issues. For the remaining eight participants, those samples were marked invalid where the corresponding pupil size was zero, as this indicated no visible pupil; for example, due to lid closure or because the pupil was outside the tracking area. For eye movements recorded in the RW, no additional calibration was required, and no further participants had to be excluded. Gaze data is generally expressed in calibrated degree visual field, with increasing values from top (gaze up) to bottom (gaze down) and left to right.

In order to relate gaze patterns with different sections of the route through the building, the continuous gaze data for both RW and VR were cut into segments according to the location of the participant along the walking route at a given time. Three different types of segments were identified for the analysis: straight walkways ("corridors"), staircases leading up ("ascending stairs") and staircases leading down ("descending stairs"). In the selected routing, the staircases leading up are interrupted by a platform with an about-turn in the middle between two floors, resulting in two stair segments per floor, whereas the descending staircases lead straight through to the next lower floor. Connecting areas and areas that could not be classified as one of the three sector types were excluded from the analysis (e. g., the turns between corridors, and a large lobby). In total, there were 6 corridors, 4 ascending stair segments and 2 descending stair segments, covering a walkway length of approximately 285 meters.

Generally, the demands of navigating the RW compared to the VR may differ, even in the most sophisticated VR model. As one possible marker of such differences, we analyzed the amount of time spent attending navigational aids, i.e., the duration the chairs with directional arrows placed among the walking route were looked at. This required us to determine the position of the chairs in the participants' field of view for each recorded frame. While in principle there exist methods in VR to conveniently identify objects hit by the observer's gaze (e.g., ray casting, Watson et al., 2019), these methods cannot be applied in the RW. To achieve comparability between the data generated from the VR and RW recordings, we chose the following method: for the real-world scene videos, we trained a deep learning algorithm (Mathis et al., 2018; Nath et al., 2019) to recognize the chairs. The algorithm delivers the 4 corner coordinates of the most likely position of a chair in each frame, together with a confidence value for its estimate. Visual inspection revealed that those positions with confidence values above 0.9 (on a scale from 0 to 1) indeed

reliably identified chair positions; positions with a confidence value below this threshold were discarded.

In VR, locating chair positions was realized by re-rendering each frame for each participant from the recorded coordinates, such that all pixels in the frame were black, except for the chairs, which were rendered blue. The identified blue pixels were then fitted with a trapezoidal shape, resulting in the 4 corner coordinates of each chair, comparable to the data obtained from the RW scene videos.

In both worlds, the distance of the current gaze from the nearest pixel contained in the chair trapezoid was then computed, and samples were considered to be on a chair whenever the distance was no larger than 1 degree. As the tracking range of the VR system is much larger than that of the RW system, chairs may be visible at distances further from the current gaze point than the maximum of the RW system. This might exaggerate the average gaze-to-chair distance in the VR world. To avoid this, the VR analysis was limited to those frames where both chair and gaze were within the corresponding tracking range of the RW system.

To visualize gaze distribution patterns in both RW and VR, heat maps were computed from gaze position data. Sample data was accumulated in 2D-histograms with a bin size of 1°, spanning a range of ±50°. Data outside this range was accumulated in the outermost bins. For display purposes, histograms were then normalized to a common range for each participant, and smoothed with a Gaussian low-pass filter (FWHH radius of 1 bin). To accommodate zero values on the logarithmic plotting scale, a regularization (+ 5% of the scale) was performed on all histograms.

The RW headset does not feature sensors for head movement recording; in VR, however, these sensors are integrated in the headset functionality as they are essential for the automatic updating of the virtual perspective. The zero-point of the headset orientation depends on the precise way in which the headset is positioned on each individual participant; we therefore chose the average position of the headset during the corridor sectors as the zero reference to allow for a comparison of head position data between sectors within the VR.

In order to improve gaze comparison between devices, we chose the visual horizon as a common point of reference for some analysis (eye-in-world). In the VR system, the horizon as well as the head angle relative to the horizon can readily be tracked. In the RW setting, however, the eye tracker used does not offer head tracking capability, and the position of the horizon in the visual field is not known. The horizon in the recorded scene video was therefore tracked by a hybrid between manual marking and a correlation-based tracking algorithm (utilizing MATLAB's xcorr2 function). Every nine frames, the horizon was marked manually in the current video frame, and the marked position was used as a reference point for the correlation tracker, which then provided the movement of the reference point for the both the following and the previous 4 frames as output. This approach for the RW scenario requires a clearly visible and identifiable horizon, at a far enough distance such that the different physical heights of the participants would not affect the angle of view at which the horizon was found in the image. One long corridor (section number 5 in **Figure 1**) with a large window at the end allowed for a reliable tracking of the far horizon and was thus chosen as the reference sector for this analysis. The vertical gaze position while passing through this sector was then subtracted from the position of the horizon on a frame-by-frame basis to achieve "eye-in-world" coordinates.

Histograms of eye-movement directions were created to profile general eye movement behavior. For each sample, the difference in gaze position relative to the previous sample was computed. Non-zero differences were then binned by direction of gaze movement, in bins of 45°, centered on the cardinal and oblique axes (resulting in a total of eight bins: [−22.5° 22.5°], [22.5° 67.5°], [67. 5° 112.5°], [112.5° 157.5°], [157.5° 202.5°], [202.5° 247.5°], [247.5° 292.5°], [292.5° 337.5°]). Histogram data was then normalized per individual to unit integral before averaging across participants.

Gaze velocity histograms were computed from gaze velocity values as defined by the absolute position difference between two neighboring valid gaze samples, normalized by the sample time difference. Gaze samples without valid neighbors were excluded from analysis. Samples were then accounted for in logarithmically spaced bins (in octaves, i.e., $< 1°$/s, 1–2°/s, 2–4°/s, 4–8°/s, 8–16°/s, 16–32°/s, 32–64°/s, 64–128°/s, 128–256°/s, 256–512°/s, $> 512°$/s).

We computed saccade rate (number of saccades per second) for each participant, in both RW and VR, separately for each sector, according to the method proposed by Engbert and Kliegl (2003), manually adjusting their algorithm's noise threshold (Lambda) individually for each participant and world.

Data was analyzed in GNU Octave (v4.4.1 and v5.2.0, Eaton et al., 2020), MATLAB (MATLAB R2019b, 2019), and R (v3.6.1, R Core Team, 2020). Repeated measures ANOVAs with factors "world" and "sector" were performed using the ezANOVA function in R (Lawrence, 2016), and Greenhouse-Geisser corrected p-values are reported along with uncorrected degrees of freedom and the Greenhouse-Geisser ε (ε_{GG}) when Mauchly's test indicated a significant violation of sphericity at a 5% level. Kolmogorov-Smirnov tests did not indicate any deviation from normality, although the sensitivity of this test (and any test for normality) is limited by the comparably low sample size.

RESULTS

For the included 8 participants (see section "Materials and Methods") data quality in the VR condition was in general better for the left than for the right eye. The left-eye data of the VR condition was thus chosen for further analysis. Sample data was mapped to degree visual field as described in the "Materials and Methods" section.

The median rendering frame rate of the VR was 73 Hz, with 95% of all frames rendered at 35 Hz or faster (this lower 5% percentile varied between 32 and 40 Hz across participants). The gaze sampling frequency of the VR tracker measured 119.1 Hz (120 Hz nominal) for 96% of all samples, with a minimum of 94% and a maximum of 99% across participants. For the VR condition, we on average recorded 400 s (SD 46 s) of data

per participant, amounting to a total of 355718 data points. Of those, 99.5% (SD 1.2%) were valid samples. Of those, all fell within the tracking window specified by the manufacturer (110° vertically and 100° horizontally, **Figure 3**). In the RW, we recorded 333 s (SD 17 s) of data, amounting to a total of 160220 data points, of which 85.5% (SD 6.6%) were valid samples, falling within the range (60° horizontally, 46° vertically) for which the manufacturer specifies tracking quality (**Figure 3B**). However, it is still reliable in which direction they are outside the tracked range (left/right and up/down). We therefore included data points outside the manufacturer-defined range in the computation of the median position and inter-quartile ranges, where their exact position does not influence these measures (given that no more than 50% of data fall outside on one side).

Real world and VR parts of the experiment did not generally last the same time (see above, paired t-test, $t(7) = -3.91$, $p = 0.006$, including the entire walking route). However, the order of the path segments was always the same as the routing through the real and virtual buildings was identical. Pairwise tests show that time spent differs significantly in the "Corridor" sectors [means: VR 167.5 s (SD 12.6), RW 123.0 s (SD 12.6), $t(7) = -8.92$, $p < 0.001$] and the "Ascending Stairs" sectors [VR 29.1 s (SD 3.2 s), RW 25.1 s (SD 1.2), $t(7) = -3.54$, $p = 0.009$], but not in the "Descending Stairs" sectors [VR 23.13 s (SD 2.5), RW 24.3 s (SD 2.5), $t(7) = 0.82$, $p = 0.439$]. In summary, participants were slower in VR for corridors and ascending stairs, but not descending stairs (**Figure 4**).

At the end of the measurement in each world, we estimated the calibration error using the same grid as used for calibration at the start for validation. This analysis revealed substantial amounts of drift (VR: 6.4° ± 5.7°, RW: 10.8° ± 2.6°) over the course of the recording, but no significant bias in drift direction for neither the VR [mean and SD, horizontal: 0.9° ± 2.0°, $t(7) = 1.35$, $p = 0.219$; vertical: −0.5° ± 8.6, $t(7) = -0.15$, $p = 0.885$] or the RW [horizontal: 0.1° ± 1.3°, $t(7) = 0.28$, $p = 0.791$; vertical: 3.3° ± 6.0°, $t(7) = 1.55$, $p = 0.166$]. No significant differences were found for the drift direction biases between VR and RW [horizontal: $t(7) = -1.15$, $p = 0.285$; vertical: $t(7) = -0.89$, $p = 0.402$]. Qualitative inspection of the data showed that within each individual all nine validation points are offset by about the same direction and magnitude, indicating that the main source of error was indeed drift, which likely resulted from movement of the headset relative to the head. Importantly, this implies that measures that are not based on absolute position – such as spread [inter-quartile-range (IQR)] and velocity – remained unaffected by this measurement error.

Gaze Distribution (Eye-in-Head)

Average eye-in-head orientation was computed separately for the three different sector types. Per participant, we characterized the gaze distribution by its median in the horizontal and vertical dimensions (**Figure 5**). Repeated measures ANOVAs with factors world (levels: VR and RW) and sector (levels: corridor, ascending stairs, and descending stairs) revealed significant main effects of vertical gaze direction for both the factor world (VR vs. RW, positive values represent downward gaze; mean of medians across participants and standard deviation: 8.59° ± 7.52° vs. 0.86° ± 9.51°, $F(1,7) = 7.34$, $p = 0.030$) and the factor sector ($F(2,14) = 33.61$, $p < 0.001$), but no significant interaction ($F(2,14) = 1.77$, $p = 0.206$). Follow-up paired t-tests (**Table 1**) show all sectors to differ from each other [corridor vs. ascending stairs, $t(7) = -3.60$, $p = 0.009$; corridor vs. descending stairs, $t(7) = -6.97$, $p < 0.001$; ascending vs. descending stairs, $t(7) = -5.21$, $p = 0.001$]. A significant main effect for the factor world was also found for horizontal gaze, although numerically the absolute difference was much smaller [−1.82° ± 2.18° vs. 2.40° ± 3.04°, $F(1,7) = 38.8$, $p < 0.001$; positive values represent rightward gaze direction]. There was no significant main effect for the factor sector on the horizontal gaze direction [$F(2,14) = 2.11$, $p = 0.158$, see **Table 1**], and no interaction [$F(2,14) = 0.56$, $p = 0.491$, $\varepsilon_{GG} = 0.54$]. These data show that the sector significantly influences gaze behavior; importantly, the lack of a significant interaction indicates that this influence is independent of whether the terrain is actually negotiated in the RW or just virtually in VR. The main effect of world in the vertical direction is somewhat surprising (if anything, one would have predicted a lower gaze in the RW). This may however be influenced by differences between the gaze recording devices or posture-related differences, as such systematic offsets are unavoidable when considering eye-in-head position data (see section "Eye-in-World. . ." below).

While the median location is a measure that is robust to outliers, in particular against points falling outside the manufacturer-specified tracking range, it is susceptible to systematic offsets and does not capture the overall distribution of the data. Consequently, we also considered a measure of spread in the horizontal and vertical dimension. The IQR is robust to both outliers (as long as outliers constitute less than 25% on either side) and offsets and thus well suited as an additional means to describe the data at hand. We computed the IQR for each participant and sector (**Figure 6**). In the horizontal direction, we found a significant main effect for the factor sector [$F(2,14) = 7.12$, $p = 0.007$], but not the factor world [$F(1,7) = 0.75$, $p = 0.415$] with no significant interaction [$F(2,14) = 0.95$, $p = 0.410$]. Similarly, in the vertical direction, we found a significant main effect for the factor sector [$F(2,14) = 9.69$, $p = 0.002$], but not the factor world [$F(1,7) = 3.23$, $p = 0.115$] with again no significant interaction [$F(2,14) = 2.68$, $p = 0.104$]. This corroborates the findings of the median position data: gaze distributions are influenced by the sector and this influence does not depend on whether the locomotion takes place in the RW or is simulated in VR.

To illustrate the individual gaze patterns visually, heat maps were generated from normalized 2D-histograms (**Figure 7**). By visual inspection, gaze patterns show substantial inter-individual differences, ranging from a narrow, focused appearance (e.g., S2) to a wide-spread pattern (e.g., S6). Participants differ both in horizontal and vertical spread. For the average of the corridor condition in the RW, an apparent two-peak pattern emerges, which is not apparent in the VR condition. Visual inspection reveals this to be due to different peak locations across individual participants rather than within. Some of the resulting distribution patterns (e.g., participant S3, **Figure 7**) resemble the T-shape previously reported in natural gaze behavior (Calow and Lappe, 2008; 't Hart et al., 2009). The T-shaped pattern is thought to

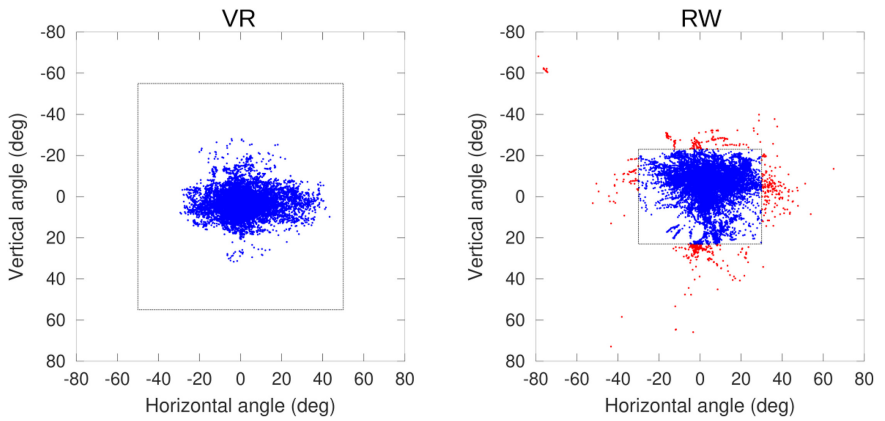


FIGURE 3 | 2D-Distribution of calibrated gaze position in head-centered coordinates, one sample participant shown. Dashed boxes indicate the manufacturer-specified tracking limits (Pupil Labs/VR: 100 × 110°, SMI/RW: 60 × 46°), red data points are outside the specified tracking limits, but the side (left/right and up/down) relative to the limits is still well-defined.

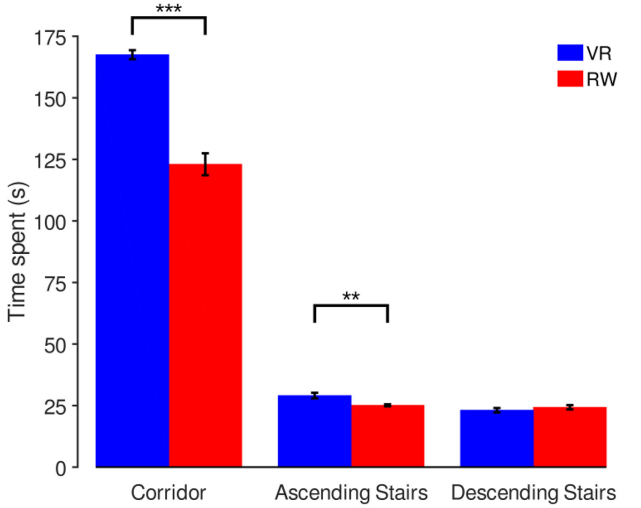


FIGURE 4 | Comparison of time spent in different sectors. Mean and SEM across participants, with significant differences indicated above (***$p < 0.001$, **$p < 0.01$).

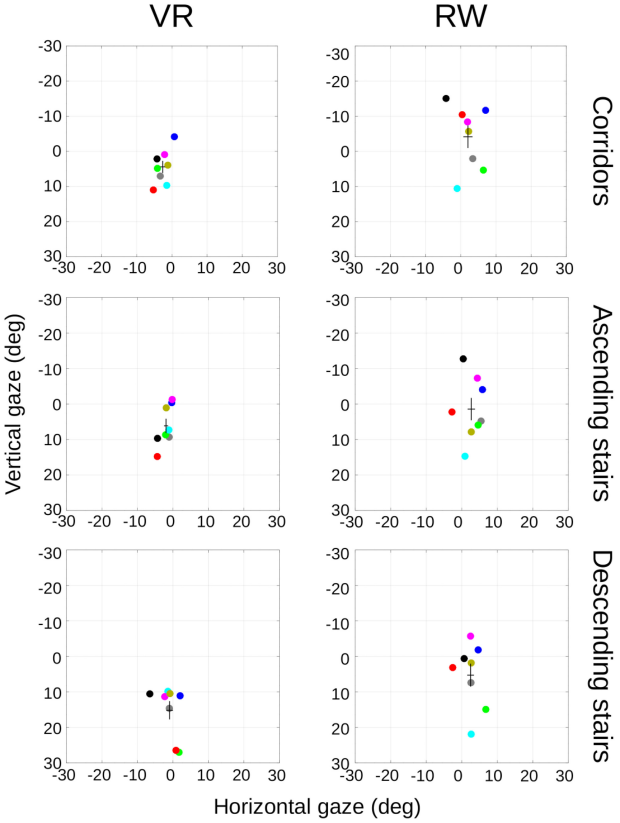


FIGURE 5 | Median of calibrated gaze positions in head-centered coordinates, separated by sector and world. Individual colored dots represent individual participants (matched across panels), black crosses represent mean and SEM across participants.

represent gaze behavior during navigation, the T-trunk resulting from gaze directed toward the terrain immediately ahead, perhaps to verify navigability, and the T-bar representing gaze directed further up and looking toward the sides, perhaps to register the surroundings or to plan further ahead. To identify possible differences in this T-shaped gaze allocation between the different worlds, we quantified the degree of T-shaped gaze distribution in each participant: the gaze data was split at the vertical median, leaving an upper and a lower half containing equally many data points. For each of these halves, the horizontal IQR was then computed, and the result of the upper half was divided by the result of the lower half. The resulting ratio was then used as input to an ANOVA with factors world and sector, as above. The difference in IQR ratios was significant only for the factor sector [$F(2,14) = 14{,}51$, $p < 0.001$], but not the factor world [$F(1,7) = 3.28$, $p = 0.110$] and there was no significant interaction [$F(2,14) = 3.29$, $p = 0.068$]. As there was a trend to

TABLE 1 | Comparison of average gaze between sectors.

Sector	Coordinate	VR	RW	*t*-score	*p*-value
Corridor	X	−2.62° ± 1.97°	2.4° ± 3.04°		
	Y	4.43° ± 4.91°	−4.15° ± 9.11°	3.11	0.017*
Ascending stairs	X	−1.92° ± 1.64°	2.69° ± 2.98°		
	Y	6.17° ± 5.71°	1.46° ± 8.92°	1.37	0.212
Descending stairs	X	−0.93° ± 2.73°	2.49° ± 2.72°		
	Y	15.16° ± 7.28°	5.27° ± 9.13°	2.73	0.029*

Mean of medians and standard deviation across participants, as well as follow-up statistics (paired t-test, VR vs. RW, df = 7) for the Y-coordinate where the factor sector had a significant main effect. Significant values ($p < 0.05$) are marked by an asterisk.

an interaction, we decided to analyze these data separately by sector. While the ratios averaged across participants were almost identical for corridors (VR: 1.16 SD 0.32; RW: 1.15 SD 0.40), the ratios on the stairs were larger in the RW, suggesting a more pronounced T-shape (ascending, VR: 1.27 SD 0.38; RW 1.58 SD 0.48; descending, VR: 1.54 SD 0.74; RW, 2.77, SD 1.42). This is an indication that for specific terrains where information from the ground is particularly relevant for foot placement (as the stairs in our case), subtle differences between VR and RW may start to emerge.

Eye-in-World: Relating Eye-in-Head Coordinates to the Horizon

The results reported so far are in head-centered (eye in head) coordinates, where position data as such may include offsets due to the different eye trackers used in the RW and VR condition. To compensate for this effect and to estimate gaze relative to the world, we computed eye-in-world coordinates by referencing gaze orientation relative to the horizon (see "Materials and Methods"). For the RW, this analysis requires the horizon to be identifiable, but at greater distance. Hence, we restricted this analysis to one corridor, where the horizon was visible through a window at the end of the hallway. No other sector shared this property, making this analysis feasible only for the chosen corridor.

On average, gaze in the VR condition was 2.2° below the horizon (SD = 7.2°) and 4.2° in the RW condition (SD = 7.0°). This difference was not significant [paired *t*-test, $t(7) = 0.55$, $p = 0.600$; **Figure 8**]. In sum, contrary to the eye-in-head data, we found no evidence for systematic differences for eye-in-world position. This makes it likely that the observed difference for eye-in-head coordinates, for which no absolute straight-ahead reference is available in VR, is primarily a consequence of headset placement relative to the participants' head. Importantly, all relative measures – spread and velocity – are insensitive to this placement as well as to its possible drift over the course of the experiment.

Head Movements

As head movement data was generally not available for the RW, we analyzed head-in-world movements in detail only for the VR. Horizontal orientation (heading) of the headset depended strongly on the position along the walking route. This stemmed on the one hand from different sectors having different compass alignments (lead heading); on the other hand, at each transition from one sector to the next, participants were physically required to turn. Due to the rectangular layout of the building, the angle of the change in route direction most often measured 90°, although the turn between segments of the upward stairs measured 180° (see **Figure 1**). Most sectors therefore start and end with a turn of at least 90°, which lead participants to make anticipating head movements in the direction of the turn as they approached the end of each sector. As the length of the individual sector types differs strongly (see **Figures 1**, **4**), head movements in the horizontal direction (yaw) for each sector type are thus contaminated to different degrees with the initiation and termination of the turns executed by the participants. We therefore limited our analysis to vertical (pitch) and roll head movements, analyzing both the mean angle and the IQR of the angular distribution in an ANOVA with the factor sector only. Average vertical head position relative to the corridors was downward 0.55° ± 2.77° for ascending stairs and downward 14.6° ± 3.64° for descending stairs; average roll relative to corridors was 0.34° ± 1.04° to the right for ascending stairs and 0.27° ± 2.08° to the left for descending stairs. We found a significant effect on average vertical head position [$F(2,14) = 86.49$, $p < 0.001$], but not on roll [$F(2,14) = 0.40$, $p = 0.681$], with no significant effect on the IQR for either vertical position [$F(2,14) = 2.86$, $p = 0.091$] or roll [$F(2,14) = 0.61$, $p = 0.559$].

Angular Distribution of Gaze Direction Changes

Eye movements during free, explorative behavior are generally highly variable. Differences in this behavior may signify differences in the processing of the visual environment. To profile these eye movements and identify possible differences between RW and VR, we assessed the directional distribution of gaze differences between recorded samples as well as the corresponding distribution of absolute gaze velocities. Similar to previous research (Einhäuser et al., 2007; Meißner et al., 2017), we find cardinal directions (horizontal/vertical) more abundant than oblique directions (**Figure 9**). To quantify this difference, we computed the ratio between the sum of the fraction of movements in cardinal directions (here defined as 45° wedges around the cardinal axes, **Figure 9**) and oblique directions and used it as input to an ANOVA with the factors world and sector. We found a significant effect for the factors world [$F(1,7) = 40.31$,

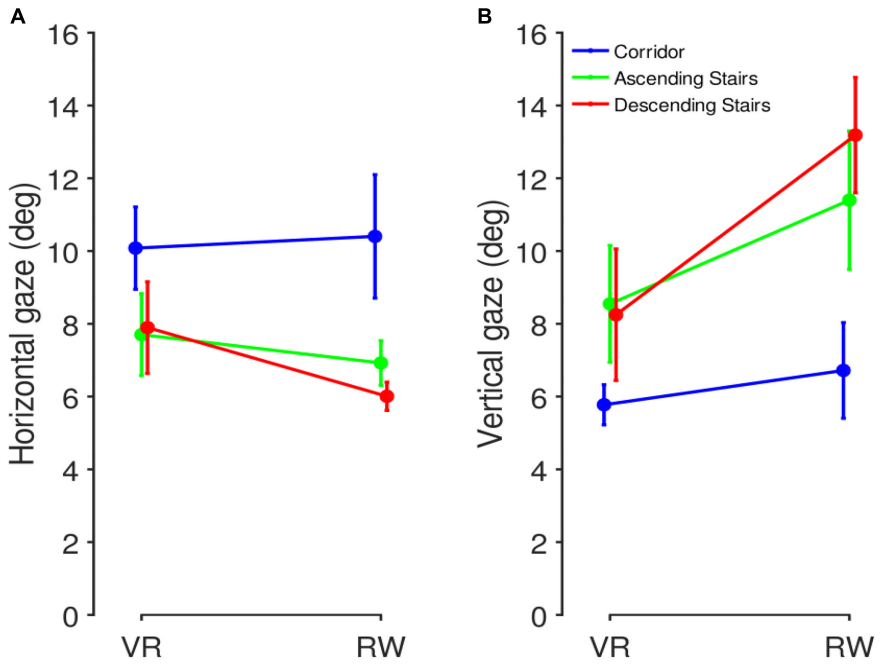


FIGURE 6 | Inter-quartile-range (IQR) comparison between RW and VR. IQRs were computed as a measure of gaze spread, separately for the three sector types. **(A)** Horizontal gaze IQR. **(B)** Vertical gaze IQR. Mean and SEM across participants.

$p < 0.001$] and sector [$F(2,14) = 5.36$, $p = 0.047$, $\varepsilon_{GG} = 0.56$], without significant interaction [$F(2,14) = 0.36$, $p = 0.706$].

Distribution of Gaze Velocities

Gaze velocities were computed and averaged across participants (**Figure 10**). The velocity distributions for RW and VR are similar in that they peak between 16 and 64°/s, but differ in that the RW measurements contained more slow velocities (<16°/s) and the VR measurements contained more fast velocities (>128°/s).

We computed an ANOVA on the per-participant medians of the velocities, with factors world and sector. We found a significant main effect for the factor world [$F(1,7) = 7.92$, $p = 0.026$], but not the factor sector [$F(2,14) = 1.89$, $p = 0.209$, $\varepsilon_{GG} = 0.55$], without significant interaction [$F(2,14) = 1.02$, $p = 0.352$, $\varepsilon_{GG} = 0.55$]. When separating this analysis by horizontal and vertical gaze component, for the horizontal component we find a significant effect for the factor world [$F(1,7) = 16.94$, $p < 0.001$] but not the factor sector [$F(2,14) = 2.87$, $p = 0.124$, $\varepsilon_{GG} = 0.60$], again with no significant interaction [$F(2,14) = 1.92$, $p = 0.184$]; We found no significant main effect or interaction for the vertical component [world: $F(1,7) = 3.26$, $p = 0.114$; sector: $F(2,14) = 1.76$, $p = 0.226$, $\varepsilon_{GG} = 0.51$; interaction: $F(2,14) = 0.47$, $p = 0.52$, $\varepsilon_{GG} = 0.51$].

Comparison of Saccade Frequency

The mean saccade rates were computed for each participant: $2.03 \pm 0.35\ s^{-1}$ for the VR (Corridors: $1.64 \pm 0.29\ s^{-1}$, Ascending stairs: $2.28 \pm 0.49\ s^{-1}$, Descending stairs: $2.21 \pm 0.70\ s^{-1}$) and $3.46 \pm 0.18\ s^{-1}$ for the RW (Corridors: $3.48 \pm 0.29\ s^{-1}$, Ascending stairs: $3.63 \pm 1.45\ s^{-1}$, Descending stairs: $3.27 \pm 1.55\ s^{-1}$). When computing an ANOVA on the per-participant saccade rates with factors world and sector, we found a significant effect for the factor world [$F(1,7) = 25.93$, $p = 0.001$], but not the factor sector [$F(2,14) = 0.91$, $p = 0.426$] with no significant interaction [$F(2,14) = 0.63$, $p = 0.545$, $\varepsilon_{GG} = 0.58$].

Dwell Time on Navigational Aids

When walking through corridors, navigation chairs were visible almost continuously, be it at the far end or nearby. However, when ascending or descending stairs, participants did not need to be directed in the proper direction due to lack of directional options. Chairs were therefore rarely within view when passing through those sectors, forcing us to limit this analysis to the corridor sectors. Average gaze dwell time on navigational chairs was 1.8% (SD = 1.4%) of the overall time spent on the walking route for the RW and 3.5% (SD = 2.3%) for VR. This difference was not significant [paired t-test, $t(7) = -1.61$, $p = 0.151$]. Average overall gaze-to-chair distance in the virtual world was 13.2° (SD = 5.8°) and in the RW 11.8° (SD = 3.3°). The difference was not significant [paired t-test, $t(7) = 0.59$, $p = 0.570$].

DISCUSSION

In the present study, we investigated how well gaze behavior in the RW can be approximated by measuring gaze in a high-fidelity VR setting. For basic measures like eye position and its spread, we found that differences between sectors (corridors, ascending stairs, and descending stairs) translated from the RW to the virtual setting, with little difference between the worlds.

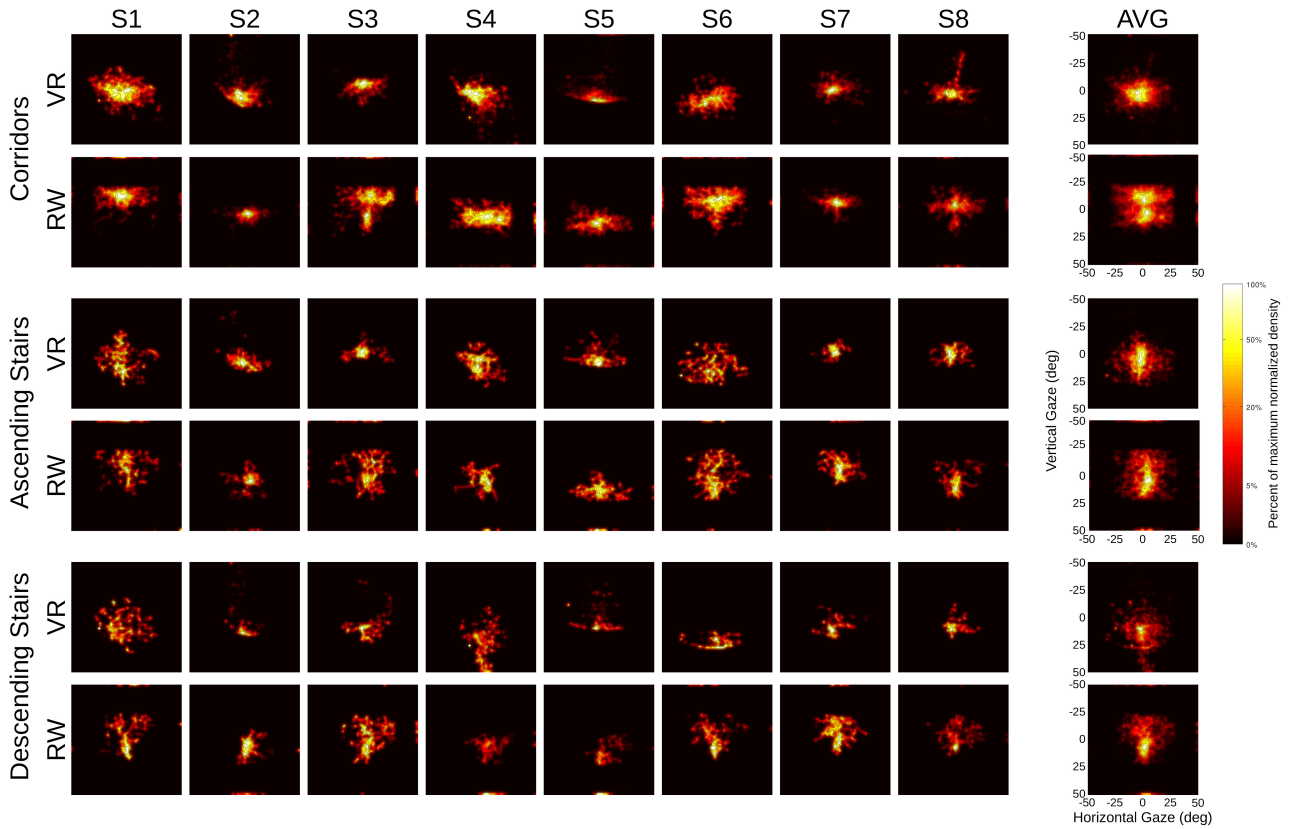


FIGURE 7 | Heat maps of gaze distribution, separated by world (virtual reality vs. real world) and sector type. Data normalized for each participant (S1 . . . S8) and displayed with a logarithmic scale (with a slight regularization to accommodate zero values), see color bar. Average data across participants in column (AVG). Outer margin in each panel represents data points outside of the specified tracking area, squares span ±50° in each dimension.

Comparison Between VR and the RW

The appearance of the simulated environment of the VR in principle cannot match the RW in all completeness. Focusing on the visual aspects of the VR employed in this study, the visual resolution of the VR system may be very high compared to previously available systems; however, it is still significantly lower than the resolution of the human visual system. The virtual copy of the chosen building, while implemented in great detail (see **Supplementary Material**), still cannot capture the richness of visual features found in the RW. As a result, a person immersed in the VR will generally be able to tell that they are not looking at the RW. The greatest benefit of the VR is the high degree of control offered by the artificial nature of the virtual surroundings. Environmental factors like the weather or third parties passing through the scenario will not affect the VR, unless desired so by the experimenter. An artificial environment has no practical size limit, and allows for arbitrary (near) real-time manipulations that would be impossible or dangerous in the RW.

Navigation in VR

In the RW, participants were required to actively walk through the setting. Natural walking behavior can support immersion in the VR (Aldaba and Moussavi, 2020; Cherep et al., 2020; Lim et al., 2020). Navigational self-localization in VR is generally enhanced if participants are allowed to move naturally while immersed in the VR (Klatzky et al., 1998; Aldaba and Moussavi, 2020). The most obvious restriction here was the need for the participants to stay physically within the range of the VR tracking range, while still promoting natural navigational behavior. However, the integration of complex VR settings with treadmills remains challenging, as it requires real-time feedback from motion capture to avoid latencies that disturb immersion. Moreover, walking in such settings is usually restricted to a small range or one linear dimension, as omnidirectional treadmills are far from widespread use as compared to the off-the-shelf head mounted display used here. Technical limitations therefore required us to keep participants within the tracking range of the VR equipment. Hence, we designed the navigation in VR as natural as possible, while participants physically remained within the tracking range, without requiring a VR cave the same size as the real-world building or a multi-directional treadmill. We exploited the observation that being able to orient the body physically appears to be important for immersion in the VR even in the absence of actual walking movement (Cherep et al., 2020; Lim et al., 2020). The solution developed here utilized the system controller, held close to the body, to orient the virtual body of the participant by orienting their

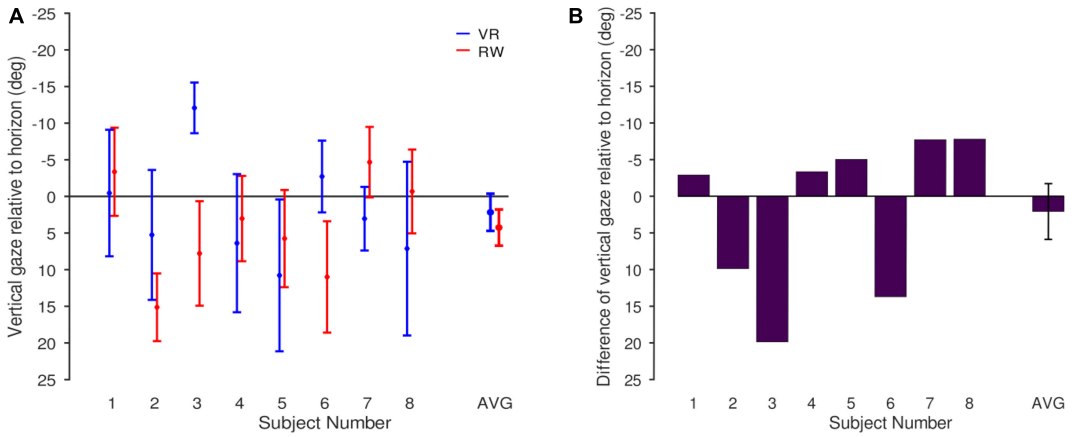


FIGURE 8 | Gaze distribution relative to the horizon for the corridor for which head-in-world direction was determined (the one depicted in **Figure 2C**, see text). **(A)** Individual participants (mean and SD) and averaged across participants (mean and SEM). **(B)** Difference between VR and RW (VR-RW) for each participant and averaged across participants (mean and SEM).

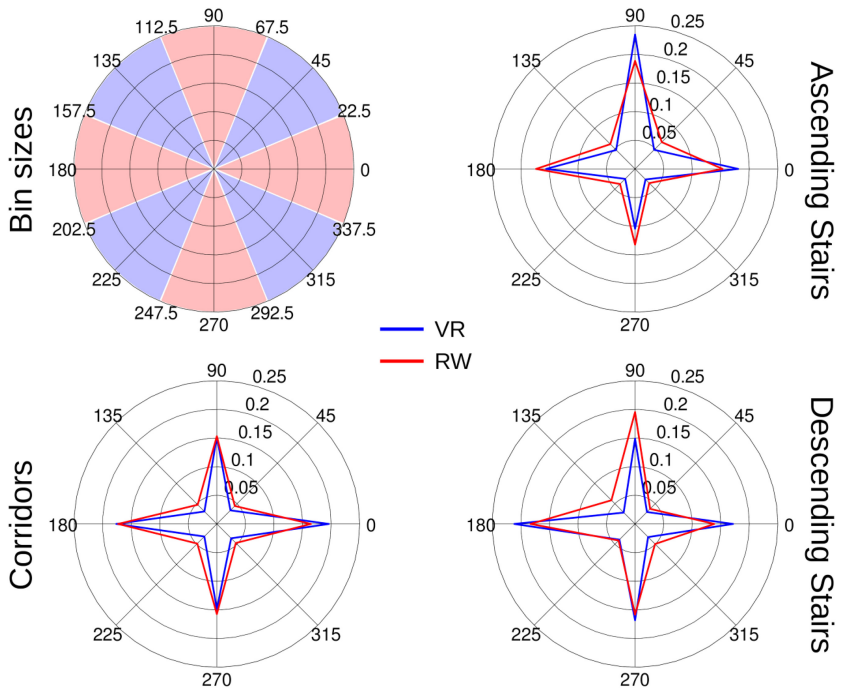


FIGURE 9 | Histograms of eye-movement directions as polar plots. Non-zero differences in gaze position between consecutive samples were accounted by direction of gaze movement in bins of 45° centered on the cardinal and oblique axes (see top–left panel). Data was normalized per individual to unit integral before averaging across participants. Directions as shown correspond to directions of gaze movements (0° corresponds to rightward, 90° to upward gaze movements, etc.).

real-world bodies in their chosen walking direction. Forward motion was controlled by pressing a button on the controller. This presents a solution to the navigation problem that minimizes the difference to the RW (Klatzky et al., 1998; Waller et al., 2004; Cherep et al., 2020; Lim et al., 2020), as orientation and navigation are still very intuitive and natural, aside from the lack of actual translational (bipedal) movement. Indeed, the analysis of the time spent looking at navigational aids (the chairs with arrows placed along the route) failed to find any significant difference between VR and RW, suggesting that there was likely no principled difference in navigational demands. The analysis of head movement data in VR shows participants moved their head down when negotiating stairs. Stairs are situations where enhanced control of foot placement would be required in the RW, but is not physically necessary in VR. The presence of said head movements is one more point suggesting that the presentation of the virtual world was convincing enough to encourage behavior that would be plausible also in the RW. The naturalness of the

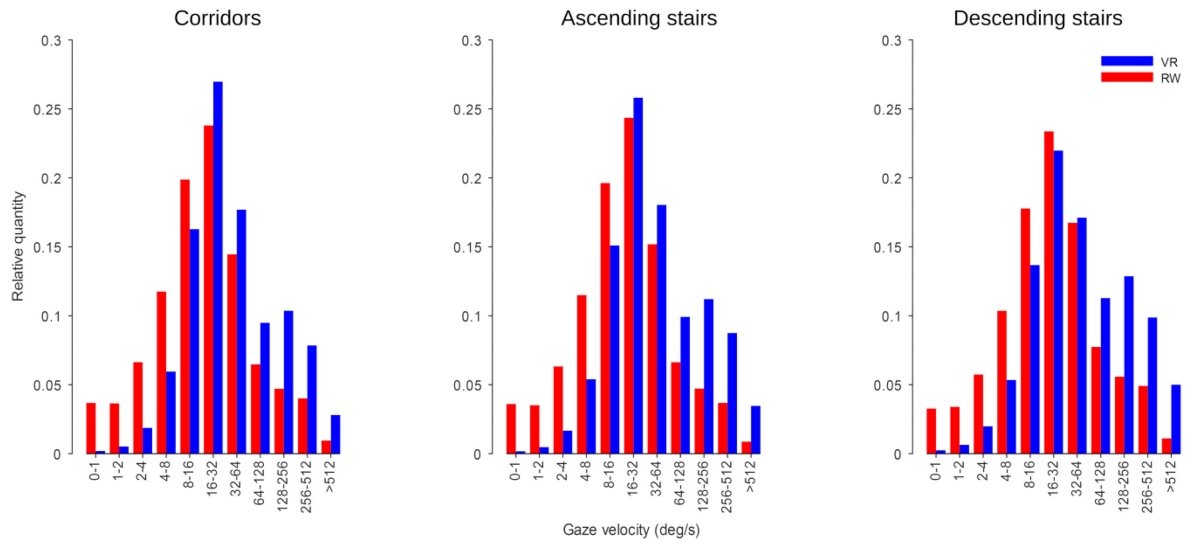


FIGURE 10 | Gaze velocity histograms. Velocity bins are log-spaced (in octaves) and labeled by the velocity contained each bin.

navigational solution may also have helped to avoid simulator sickness (no incidences were reported by our participants), which can otherwise be a problem when moving in virtual realities (Golding and Gresty, 2005; Dużmańska et al., 2018).

A simple extension to this approach would be to physically attach the controller to the body of the participant, possible at the hip, which would free one hand and allow for an even more natural posture during exposure to the VR. Our participants did not report any subjective difficulties with the employed method of navigation, and there was no occurrence of cyber sickness. The usage of the standard VR controller for this purpose helps not only to reduce the cost of acquisition of the VR setup, but also cuts down the complexity of the required software development, which will facilitate further experimentation in the future. Generally, while current methods of navigating a virtual environment differ in many cases from the natural means of bipedal movement, this may also offer new chances and opportunities, e.g., in medical rehabilitation training, where patients may be unable to execute the full range of movements available to healthy controls.

Gaze in VR and RW

Drift – i.e., a growing offset between actual and measured gaze direction that applies uniformly to the whole measured field – is a significant and well-documented factor in head-mounted eye tracking equipment (Sugano and Bulling, 2015; Müller et al., 2019); indeed, the absolute drift in our experiments was quite substantial as compared to stationary eye-tracking equipment. However, there was no significant bias in drift direction for either VR or RW, as well as no significant difference between the RW and VR. This suggests that the results presented here were not systematically affected by changes in position of the measurement equipment during the course of the experiment. Moreover, all measures but the eye-in-head direction, are by construction insensitive against these drifts. Where we *did* consider eye-in-head directions, especially in the gaze-distribution maps of figure 7, the sizes of the observed patterns were large compared to the effects of drift, such that drift is unlikely to have affected these patterns qualitatively. This also applies to individual differences among these gaze maps, which are substantial, a pattern consistent with previous observations on natural scene viewing (e.g., Yarbus, 1967; de Haas et al., 2019).

When real-world gaze allocation is compared to standard laboratory eye-tracking settings, profound differences are found, in particular with respect to gaze in direction of the ground ('t Hart et al., 2009). However, there are multiple differences between walking through the RW and watching the same visual input on a screen: the visual input on the screen is limited in visual field and resolution, head and body are restrained and there is no need to actively navigate or walk through the environment. To isolate the component of safe walking from the other differences, we here attempted to approximate the natural situation with respect to its visual appearance and its navigational requirements as closely as technically possible in VR. As expected from real-world studies ('t Hart and Einhäuser, 2012), we found profound differences between different terrains (sectors) for nearly all of the measures tested. One might have also expected differences between the worlds or an interaction of the world with terrain (if the VR had been perceived as entirely unconvincing by the participants, gaze patterns in VR might have differed less between the different sectors than in RW). In particular, one might assume that the additional requirement to place one's feet carefully in the RW as compared to VR (Matthis et al., 2018; Kopiske et al., 2020; Thomas et al., 2020) would be accompanied by significant changes in gaze behavior, especially when negotiating the stairs. Surprisingly, however, we found no interactions between the factors world and sector for any of the measures tested. Effects of the world were found for the vertical gaze direction in eye-in-head coordinates, the vertical

head orientation in VR, the number of saccades made and a subtle difference in the angular distribution of gaze-direction changes. The direction of the former effect – gaze was lower in VR on average than in the RW – was contrary to expectations ('t Hart et al., 2009; 't Hart and Einhäuser, 2012): one would expect virtual locomotion to require fewer looks to the ground where the information for foot placement is gathered in the RW (Marigold and Patla, 2007) and also during actual walking in VR (Kopiske et al., 2020). However, this effect is likely explained by headset placement and absent (numerically even reversed) when gaze is referenced to the horizon. As a measure that is insensitive to offsets in the headset placement, we quantified the spread of eye movements by using the IQR (**Figure 6**). As before, we found significant differences only between the sectors, not between the worlds, and importantly no interaction between the factors. This underlines the observation that the differences between sectors translate well from the RW to the VR, and – for our setting – differences between the worlds are minute. The differences in angular distribution of gaze-direction changes between the worlds are also subtle, provided the comparably large differences found between different real-world environments (Einhäuser et al., 2007), which in turn are comparable to the differences between sectors in the present study. It is tempting to speculate that similar factors generate the differences and are related to the requirements of the environment, with more liberal (less navigation-driven) exploration in the VR generating more cardinal eye movements. Generally, fixations and smooth pursuit are not trivial to tell apart in head-free scenarios, as what looks like pursuit in gaze angle velocities may indeed be a fixation on a physically stationary object in the presence of head movements. Additionally, a physically stationary object such as a light switch on a wall may move through the observer's field of view on a path consistent with the optical flow as the observer moves forward. The gaze velocity analysis indicates relatively more saccades in the RW, suggesting more exploratory saccades, perhaps due to a more navigation-driven exploratory behavior. This is also supported by the relative increase of higher velocities in the gaze velocity analysis, which in turn finds more slow eye movements in the VR. This would be well explained by an increased number of optic flow linked fixations as a counterpoint to the increased number of saccades in RW.

For the "T-shape" previously described in the RW (Calow and Lappe, 2008; 't Hart et al., 2009), we find a trend to an interaction between world and sector, so we cannot exclude that differences between VR and RW will start to emerge when more sophisticated measures or more difficult terrain (as compared to the smooth floor surface of an office building) are concerned. Explicitly modeling difficult and irregular terrain in VR will therefore become an interesting line for further research (cf. Kopiske et al., 2020).

CONCLUSION

In summary, we found surprisingly little difference between gaze behavior in VR and RW for our setting; to the contrary, virtual locomotion seems to capture the major differences between different environmental constraints (the factor "sector" in our experiment) remarkably well. The effects of world (VR vs. RW) we found were either well-explainable by equipment particularities, as for the vertical eye-in-head position, or subtle compared to previously reported differences between different real-world settings, as in the case of the cardinal preferences. This opens up an avenue of possibility for research that would previously have been possible only in real-world settings, but with the enhanced control over environmental factors offered by VR that would otherwise be largely left at random. Gaze analysis in life-like settings, but still under highly controlled conditions, has therefore now become a tangible reality. Remaining factors that may affect the depth of immersion and thus also the similarity of the gaze behavior in simulated environments may be addressed through improved VR devices, such as treadmills to allow for even more realistic navigation (Kopiske et al., 2020) or improvements in available computational power for even more visual details.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

All procedures were evaluated by the applicable ethics board (Ethikkommission der Fakultät HSW, TU Chemnitz) who ruled that no in-depth evaluation was necessary (case-no.: V-274-PHKP-WET-Augenb-11062018). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors conceived the study, designed the experiment, and reviewed the manuscript. SF created the virtual reality model. SF and WE collected the data. JD and SF analyzed the data. JD and WE wrote the manuscript.

ACKNOWLEDGMENTS

We thank Philipp Methfessel for supporting the deep-learning based annotation, as well as Elisa-Maria Heinrich, Christiane Breitkreutz, and Fabian Parth for annotating the real-world videos and Alexandra Bendixen for feedback on experimental design and data analysis. The work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 222641018 – SFB/TRR 135 TP B1 (WE) – and project ID 416228727 – SFB 1410 TP A04 (SF).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.656913/full#supplementary-material

REFERENCES

Aldaba, C. N., and Moussavi, Z. (2020). Effects of virtual reality technology locomotive multi-sensory motion stimuli on a user simulator sickness and controller intuitiveness during a navigation task. *Med. Biol. Eng. Comput.* 58, 143–154. doi: 10.1007/s11517-019-02070-2

Anderson, N. C., Bischof, W. F., Foulsham, T., and Kingstone, A. (2020). Turning the (virtual) world around: patterns in saccade direction vary with picture orientation and shape in virtual reality. *PsyArXiv [Preprints]*

Baheux, K., Yoshizawa, M., Tanaka, A., Seki, K., and Handa, Y. (2005). Diagnosis and rehabilitation of hemispatial neglect patients with virtual reality technology. *Technol. Health Care* 13, 245–260.

Bernardo, A. (2017). Virtual reality and simulation in neurosurgical training. *World Neurosurg.* 106, 1015–1029. doi: 10.1016/j.wneu.2017.06.140

Bruce, N. D. B., and Tsotsos, J. K. (2006). A statistical basis for visual field anisotropies. *Neurocomputing* 69, 1301–1304. doi: 10.1016/j.neucom.2005.12.096

Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology and Perception in Art*. Oxford: Univ. Chicago Press.

Calow, D., and Lappe, M. (2008). Efficient encoding of natural optic flow. *Network* 19, 183–212. doi: 10.1080/09548980802368764

Chapman, P. R., and Underwood, G. (1998). Visual search of driving situations: danger and experience. *Perception* 27, 951–964.

Cherep, L. A., Lim, A. F., Kelly, J. W., Acharya, D., Velasco, A., Bustamante, E., et al. (2020). Spatial cognitive implications of teleporting through virtual environments. *J. Exp. Psychol. Appl.* 26, 480–492. doi: 10.1037/xap0000263

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., and Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *PNAS* 116, 11687–11692. doi: 10.1073/pnas.1820553116

Drewes, J., Zhu, W., Hu, Y., and Hu, X. (2014). Smaller is better: drift in gaze measurements due to pupil dynamics. *PLoS One* 9:e111197. doi: 10.1371/journal.pone.0111197

Dużmańska, N., Strojny, P., and Strojny, A. (2018). Can simulator sickness be avoided? a review on temporal aspects of simulator sickness. *Front. Psychol.* 9:2132. doi: 10.3389/fpsyg.2018.02132

Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. (2020). *GNU Octave Version 5.2.0 Manual: A High-Level Interactive Language for Numerical Computations*. Available online at: https://www.gnu.org/software/octave/doc/v5.2.0/ (accessed April 9, 2021).

Einhäuser, W., Rutishauser, U., and Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *J. Vision* 8:2. doi: 10.1167/8.2.2

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., et al. (2007). Human eye-head co-ordination in natural exploration. *Network* 18, 267–297. doi: 10.1080/09548980701671094

Engbert, R., and Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Res.* 43, 1035–1045. doi: 10.1016/S0042-6989(03)00084-1

Fetter, M. (2007). Vestibulo-ocular reflex. *Dev. Ophthalmol.* 40, 35–51. doi: 10.1159/000100348

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Res.* 51, 1920–1931. doi: 10.1016/j.visres.2011.07.002

Garbutt, S., Harwood, M. R., and Harris, C. M. (2001). Comparison of the main sequence of reflexive saccades and the quick phases of optokinetic nystagmus. *Br. J. Ophthalmol.* 85, 1477–1483. doi: 10.1136/bjo.85.12.1477

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., and Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. *J. Vision* 12, 17–17. doi: 10.1167/12.6.17

Georgescu, A. L., Kuzmanovic, B., Roth, D., Bente, G., and Vogeley, K. (2014). The use of virtual characters to assess and train non-verbal communication in high-functioning autism. *Front. Hum. Neurosci.* 8:807. doi: 10.3389/fnhum.2014.00807

Golding, J. F., and Gresty, M. A. (2005). Motion sickness. *Curr. Opin. Neurol.* 18, 29–34. doi: 10.1097/00019052-200502000-00007

Harel, J., Koch, C., and Perona, P. (2006). "Graph-based visual saliency," in *Proceedings of the 19th International Conference on Neural Information Processing Systems NIPS'06*, (Cambridge: MIT Press), 545–552.

Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194. doi: 10.1016/j.tics.2005.02.009

Hayhoe, M. M., McKinney, T., Chajka, K., and Pelz, J. B. (2012). Predictive eye movements in natural vision. *Exp. Brain Res.* 217, 125–136. doi: 10.1007/s00221-011-2979-2

Helbing, J., Draschkow, D., and Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition* 196:104147. doi: 10.1016/j.cognition.2019.104147

Henderson, J. M., and Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 1, 743–747. doi: 10.1038/s41562-017-0208-0

Henderson, J. M., Larson, C. L., and Zhu, D. C. (2007). Cortical activation to indoor versus outdoor scenes: an fMRI study. *Exp. Brain Res.* 179, 75–84. doi: 10.1007/s00221-006-0766-2

Hong, Y.-J., Kim, H. E., Jung, Y. H., Kyeong, S., and Kim, J.-J. (2017). Usefulness of the mobile virtual reality self-training for overcoming a fear of heights. *Cyberpsychol. Behav. Soc. Netw.* 20, 753–761. doi: 10.1089/cyber.2017.0085

Huang, T.-K., Yang, C.-H., Hsieh, Y.-H., Wang, J.-C., and Hung, C.-C. (2018). Augmented reality (AR) and virtual reality (VR) applied in dentistry. *Kaohsiung J. Med. Sci.* 34, 243–248. doi: 10.1016/j.kjms.2018.01.009

Ilg, U. J. (1997). Slow eye movements. *Prog. Neurobiol.* 53, 293–329.

Ilg, U. J. (2002). "Commentary: smooth pursuit eye movements: from low-level to high-level vision," in *Progress in Brain Research*, Vol. 140, eds J. Hyona, D. P. Munoz, W. Heide, and R. Radach (Amsterdam: Elsevier), 279–298. doi: 10.1016/S0079-6123(02)40057-X

Iskander, J., Hossny, M., and Nahavandi, S. (2019). Using biomechanics to investigate the effect of VR on eye vergence system. *Appl. Ergon.* 81:102883. doi: 10.1016/j.apergo.2019.102883

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Machine Intellig.* 20, 1254–1259. doi: 10.1109/34.730558

Izard, S. G., Juanes, J. A., García Peñalvo, F. J., Estella, J. M. G., Ledesma, M. J. S., and Ruisoto, P. (2018). Virtual reality as an educational and training tool for medicine. *J. Med. Syst.* 42:50. doi: 10.1007/s10916-018-0900-2

Jones, P. R., Somoskeöy, T., Chow-Wing-Bom, H., and Crabb, D. P. (2020). Seeing other perspectives: evaluating the use of virtual and augmented reality to simulate visual impairments (OpenVisSim). *NPJ Digit. Med.* 3:32. doi: 10.1038/s41746-020-0242-6

Kapitaniak, B., Walczak, M., Kosobudzki, M., Jóźwiak, Z., and Bortkiewicz, A. (2015). Application of eye-tracking in the testing of drivers: a review of research. *Int. J. Occup. Med. Environ. Health* 28, 941–954. doi: 10.13075/ijomeh.1896.00317

Kim, H., Shin, J. E., Hong, Y.-J., Shin, Y.-B., Shin, Y. S., Han, K., et al. (2018). Aversive eye gaze during a speech in virtual environment in patients with social anxiety disorder. *Aust. N. Z. J. Psychiatry* 52, 279–285. doi: 10.1177/0004867417714335

Klatzky, R. L., Loomis, J. M., Beall, A. C., Chance, S. S., and Golledge, R. G. (1998). Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychol. Sci.* 9, 293–298.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Konstantopoulos, P., Chapman, P., and Crundall, D. (2010). Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers' eye movements in day, night and rain driving. *Accid Anal. Prev.* 42, 827–834. doi: 10.1016/j.aap.2009.09.022

Kopiske, K. K., Koska, D., Baumann, T., Maiwald, C., and Einhäuser, W. (2020). Icy road ahead - rapid adjustments of gaze-gait interactions during perturbed naturalistic walking. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/mabn4

Kowler, E. (2011). Eye movements: the past 25years. *Vision Res.* 51, 1457–1483. doi: 10.1016/j.visres.2010.12.014

Kramida, G. (2016). Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Trans. Vis. Comput. Graph.* 22, 1912–1931. doi: 10.1109/TVCG.2015.2473855

Kredel, R., Vater, C., Klostermann, A., and Hossner, E.-J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: a systematic review of 40 years of research. *Front. Psychol.* 8:1845. doi: 10.3389/fpsyg.2017.01845

Kümmerer, M., Theis, L., and Bethge, M. (2015). *Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on Imagenet*. San Diego, CA: ICLR.

Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., and Sarkar, N. (2015). A physiologically informed virtual reality based social communication system for individuals with autism. *J. Autism. Dev. Disord.* 45, 919–931. doi: 10.1007/s10803-014-2240-5

Land, M. F. (1992). Predictable eye-head coordination during driving. *Nature* 359, 318–320. doi: 10.1038/359318a0

Land, M. F., and McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nat. Neurosci.* 3, 1340–1345. doi: 10.1038/81887

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. Available online at: https://CRAN.R-project.org/package=ez (accessed April 9, 2021).

Lim, A., Kelly, J., Sepich, N., Cherep, L., Freed, G., and Gilbert, S. (2020). "Rotational self-motion cues improve spatial learning when teleporting in virtual environments," in *Proceedings of the SUI '20: Symposium on Spatial User Interaction*, (New York, NY: ACM, Inc), doi: 10.1145/3385959.3418443

Magnusson, M., Pyykkö, I., and Norrving, B. (1986). The relationship of optokinetic nystagmus to pursuit eye movements, vestibular nystagmus and to saccades in humans. A clinical study. *Acta Oto Laryngol.* 101, 361–370. doi: 10.3109/00016488609108620

Marigold, D. S., and Patla, A. E. (2007). Gaze fixation patterns for negotiating complex ground terrain. *Neuroscience* 144, 302–313. doi: 10.1016/j.neuroscience.2006.09.006

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* 5, 229–240. doi: 10.1038/nrn1348

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., et al. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21, 1281–1289.

MATLAB R2019b (2019). *MathWorks Announces Release 2019b of MATLAB and Simulink*. Natick, MA: The MathWorks Inc.

Matthis, J. S., Yates, J. L., and Hayhoe, M. M. (2018). Gaze and the control of foot placement when walking in natural terrain. *Curr. Biol.* 28, 1224–1233.e5. doi: 10.1016/j.cub.2018.03.008

Mazloumi Gavgani, A., Walker, F. R., Hodgson, D. M., and Nalivaiko, E. (2018). A comparative study of cybersickness during exposure to virtual reality and "classic" motion sickness: are they different? *J. Appl. Physiol. (1985)* doi: 10.1152/japplphysiol.00338.2018 [Epub ahead of print].

Meißner, M., Pfeiffer, J., Pfeiffer, T., and Oppewal, H. (2017). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. *J. Business Res.* 100, 445–458. doi: 10.1016/j.jbusres.2017.09.028

Mon-Williams, M., Plooy, A., Burgess-Limerick, R., and Wann, J. (1998). Gaze angle: a possible mechanism of visual stress in virtual reality headsets. *Ergonomics* 41, 280–285. doi: 10.1080/001401398187035

Müller, P., Buschek, D., Huang, M. X., and Bulling, A. (2019). "Reducing calibration drift in mobile eye trackers by exploiting mobile phone usage," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research and Applications, 1–9*, (New York, NY: ACM, Inc), doi: 10.1145/3314111

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., and Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* 14, 2152–2176. doi: 10.1038/s41596-019-0176-0

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rolfs, M. (2009). Microsaccades: small steps on a long way. *Vision Res.* 49, 2415–2441. doi: 10.1016/j.visres.2009.08.010

Rothkopf, C. A., and Ballard, D. H. (2009). Image statistics at the point of gaze during human navigation. *Vis. Neurosci.* 26, 81–92. doi: 10.1017/S0952523808080978

Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2007). Task and context determine where you look. *J. Vision* 7, 16–16. doi: 10.1167/7.14.16

Spering, M., and Montagnini, A. (2011). Do we track what we see? Common versus independent processing for motion perception and smooth pursuit eye movements: a review. *Vision Res.* 51, 836–852. doi: 10.1016/j.visres.2010.10.17

Steinman, R. M., and Collewijn, H. (1980). Binocular retinal image motion during active head rotation. *Vision Res.* 20, 415–429. doi: 10.1016/0042-6989(80)90032-2

Stoll, J., Thrun, M., Nuthmann, A., and Einhäuser, W. (2015). Overt attention in natural scenes: objects dominate features. *Vision Res.* 107, 36–48. doi: 10.1016/j.visres.2014.11.006

Sugano, Y., and Bulling, A. (2015). "Self-calibrating head-mounted eye trackers using egocentric visual saliency," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, (New York, NY), 363–372. doi: 10.1145/2807442.2807445

Swan, G., Goldstein, R. B., Savage, S. W., Zhang, L., Ahmadi, A., and Bowers, A. R. (2020). Automatic processing of gaze movements to quantify gaze scanning behaviors in a driving simulator. *Behav. Res.* 53, 487–506. doi: 10.3758/s13428-020-01427-y

't Hart, B. M., and Einhäuser, W. (2012). Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Exp. Brain Res.* 223, 233–249. doi: 10.1007/s00221-012-3254-x

't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., et al. (2009). Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. *Visual Cogn.* 17, 1132–1158. doi: 10.1080/13506280902812304

Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *J. Vision* 11:5. doi: 10.1167/11.5.5

Thoma, V., and Dodd, J. (2019). "Web usability and eyetracking," in *Eye Movement Research: An Introduction to its Scientific Foundations and Applications Studies in Neuroscience, Psychology and Behavioral Economics*, eds C. Klein and U. Ettinger (Cham: Springer International Publishing), 883–927. doi: 10.1007/978-3-030-20085-5_21

Thomas, N. D. A., Gardiner, J. D., Crompton, R. H., and Lawson, R. (2020). Physical and perceptual measures of walking surface complexity strongly predict gait and gaze behaviour. *Hum. Movement Sci.* 71:102615. doi: 10.1016/j.humov.2020.102615

Turnbull, P. R. K., and Phillips, J. R. (2017). Ocular effects of virtual reality headset wear in young adults. *Sci. Rep.* 7:16172. doi: 10.1038/s41598-017-16320-6

Underwood, G., and Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Q. J. Exp. Psychol. (Hove)* 59, 1931–1949. doi: 10.1080/17470210500416342

van der Veer, A., Alsmith, A., Longo, M., Wong, H. Y., Diers, D., Bues, M., et al. (2019). "The Influence of the Viewpoint in a Self-Avatar on Body Part and Self-Localization," in *Peoceedings of the ACM Symposium on Applied Perception 2019 SAP '19*, (New York, NY: Association for Computing Machinery), 1–11. doi: 10.1145/3343036.3343124

Waller, D., Loomis, J. M., and Haun, D. B. M. (2004). Body-based senses enhance knowledge of directions in large-scale environments. *Psychonomic Bull. Rev.* 11, 157–163. doi: 10.3758/BF03206476

Watson, M. R., Voloh, B., Thomas, C., Hasan, A., and Womelsdorf, T. (2019). USE: an integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. *J. Neurosci. Methods* 326:108374. doi: 10.1016/j.jneumeth.2019.108374

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). SUN: a Bayesian framework for saliency using natural statistics. *J. Vision* 8, 32.1–20. doi: 10.1167/8.7.32

Zhang, L., Wade, J., Bian, D., Fan, J., Swanson, A., Weitlauf, A., et al. (2017). Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE Trans. Affect. Comput.* 8, 176–189. doi: 10.1109/TAFFC.2016.2582490

Zhang, S., McClean, S. I., Garifullina, A., Kegel, I., Lightbody, G., Milliken, M., et al. (2018). "Evaluation of the TV customer experience using eye tracking technology. in BCS, the chartered institute for IT," in *Proceedings of the 32nd Human Computer Interaction Conference (British Computer Society)*, (Swindon), doi: 10.14236/ewic/HCI2018.88

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fnins.2021.664490

Imaging Time Series of Eye Tracking Data to Classify Attentional States

Lisa-Marie Vortmann[1]*, *Jannes Knychalla*[1], *Sonja Annerer-Walcher*[2], *Mathias Benedek*[2] and *Felix Putze*[1]

[1] Cognitive Systems Lab, Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany,
[2] Creative Cognition Lab, Institute of Psychology, University of Graz, Graz, Austria

It has been shown that conclusions about the human mental state can be drawn from eye gaze behavior by several previous studies. For this reason, eye tracking recordings are suitable as input data for attentional state classifiers. In current state-of-the-art studies, the extracted eye tracking feature set usually consists of descriptive statistics about specific eye movement characteristics (i.e., fixations, saccades, blinks, vergence, and pupil dilation). We suggest an Imaging Time Series approach for eye tracking data followed by classification using a convolutional neural net to improve the classification accuracy. We compared multiple algorithms that used the one-dimensional statistical summary feature set as input with two different implementations of the newly suggested method for three different data sets that target different aspects of attention. The results show that our two-dimensional image features with the convolutional neural net outperform the classical classifiers for most analyses, especially regarding generalization over participants and tasks. We conclude that current attentional state classifiers that are based on eye tracking can be optimized by adjusting the feature set while requiring less feature engineering and our future work will focus on a more detailed and suited investigation of this approach for other scenarios and data sets.

Keywords: convolutional neural network, eye tracking, classification, Imaging Time Series, Augmented Reality, Gramian Angular Fields, Markov Transition Fields, attention

OPEN ACCESS

Edited by:
Julien Epps,
University of New South Wales, Australia

Reviewed by:
Jane Zhen Liang,
Shenzhen University, China
Hong Zeng,
Southeast University, China
Markku Tukiainen,
University of Eastern Finland, Finland
Thomas Kuebler,
University of Tübingen, Germany

***Correspondence:**
Lisa-Marie Vortmann
vortmann@uni-bremen.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *05 February 2021*
Accepted: *03 May 2021*
Published: *28 May 2021*

Citation:
Vortmann L-M, Knychalla J, Annerer-Walcher S, Benedek M and Putze F (2021) Imaging Time Series of Eye Tracking Data to Classify Attentional States. Front. Neurosci. 15:664490. doi: 10.3389/fnins.2021.664490

1. INTRODUCTION

Scientists' fascination for human eye gaze behavior started as early as in the 19th century when it was observed that the eyes don't move in one fluent motion while reading. Instead, they stop and focus often but only briefly. This observation led to many questions: When do they stop? Where do they focus and how long? And most importantly, why? In 1908, Edmund Huey published the first version of his book "The psychology and pedagogy of reading" (Huey, 1908) in which he discussed these observations and introduced one of the first versions of an eye tracking device. It consisted of a special contact lens that was connected to an aluminum pointer. Since then, the field of eye tracking has flourished and continuously improved eye tracking devices. In 1980, Marcel Adam Just and Patricia A. Carpenter proposed their Eye-Mind assumption, stating that "there is no appreciable lag between what is being fixated and what is being processed" (Just and Carpenter, 1980). While, this statement is restricted to eye fixations, it can be assumed that gaze behavior, in general, is closely tied to mental processes. Our knowledge about saccades and fixations, their cause and reason, and their connection to the current mental state of the observed person has increased immensely since then and the practice of eye tracking has found many applications. In addition to the mentioned

research interests, human gaze tracking is widely used in consumer and marketing research (Wedel and Pieters, 2008) or as an input mechanism for technical devices, such as smartphones (Paletta et al., 2014) and Augmented and Virtual Reality glasses (Miller, 2020).

Some applications are mainly interested in the direction of the gaze (i.e., to predict salient regions of web pages as in Buscher et al., 2009). Others, however, make use of implications about the mental state that can be drawn from the eye tracking data. One famous and possibly life-saving use of eye tracking is to detect a high cognitive workload (Palinko et al., 2010), or high level of fatigue (Horng et al., 2004) in car drivers. Di Stasi et al. (2013) suggested that ocular instability increases with mental fatigue, meaning that saccadic and microsaccadic velocity decreases and drift velocity increases. If this movement behavior is observed in a driver, they can be advised to take a break from driving.

Another interesting application field for mental state classification that is gaining interest in the current Covid-19 pandemic is digital learning settings. The learning system could for example detect phases of mind-wandering. This information about the mental state of the learner can then be used to later present the corresponding content again during phases of concentration and thus, improve the chances of a better learning rate and greater learning success (Conati et al., 2013). The aspects of the human mental state that can be classified or detected are manifold. Besides the mentioned workload, fatigue, and mind-wandering, further cognitive and affective states can be modeled, such as internally and externally directed attention, attentional shifts, emotions, the direction of attention, goal-directed and task-related internal attention, or alertness.

In many studies, mental state classification is based on data from other biosignals, such as brain activity. Often, electroencephalography (EEG) is chosen for its good temporal resolution and low cost (in comparison to fMRI), as for example in Zeng et al. (2018), Dehais et al. (2018), Vézard et al. (2015), Benedek et al. (2014), Ceh et al. (2020), and Vortmann et al. (2019a). However, compared to eye tracking devices, the setup time highly depends on the number of electrodes and usually requires qualified assistance for the user. In comparison, eye tracking has the obvious advantages of a fast setup, easy calibration, and the fact that eye tracking glasses promise a better usability experience in the wild than tight EEG-caps.

The movement of the eyes is typically recorded as a time series of gaze point coordinates from both eyes. Some systems additionally record pupil diameters or blinks. Once this data is acquired, it needs to be processed so the important information can be extracted and used to draw conclusions about the mental state of the user. Typical features that are calculated on the data include the number and length of fixations, saccades, and microsaccades, the gaze velocity, the pupil size, the frequency of blinks, or the covered gaze distance. With this set of features, a supervised machine learning algorithm can learn to model the mental states of interest and detect these states in the user. One major challenge in improving the accuracy of mental state classification based on eye tracking data is finding and optimizing the right features and algorithms. In recent years, the machine learning community has solved more and more problems using deep learning approaches and neural nets because they require less feature engineering and are thus more suitable if there is a lack of domain understanding. They are used in a variety of scenarios from forecasting to fraud detection and financial services or image recognition.

Wang and Oates (2015) suggested that time series data could be represented as images or matrices (Imaging Time Series, ITS) and then these can be classified by Convolutional Neural Networks (CNN) which have proven to be successful in image classification in the past. To transform the variables from one-dimensional time series to two-dimensional images, they suggest two different algorithms: Gramian Angular Fields (GAF) which represent the temporal correlation between time points, and Markov Transition Fields (MTF) which calculate a matrix based on transition probabilities (see section 2.2.2).

In this work, we compare one-dimensional (1D) statistical summary feature set based approaches with ITS approaches for the detection of attentional states on three different eye tracking data sets related to attention. The first data set contains phases of internally and externally directed attention during several screen-based tasks (see section 2.1.1). The second data set is on the same aspect of attention but was collected in an Augmented Reality scenario (see section 2.1.2). Likewise, the third data set was collected during an Augmented Reality task but consists of phases on attention on real and phases of attention on virtual objects (see section 2.1.3). The aim is to improve the classification accuracy for multiple aspects of attention for both person-dependently and person-independently trained models. To the best of our knowledge, no previous study has performed such a comparison with the suggested methods on eye tracking data.

1.1. Related Work on Mental State Classification From Eye Behavior

Related studies that aimed at classifying mental states and especially attentional states from eye tracking data guided us in finding state-of-the-art features for our 1D statistical summary feature set and gave us an overview over which algorithms should be used for the comparison. Additionally, their results show that it is possible to reliably detect these states in eye tracking data.

The popular topic of eye movements during reading tasks was picked up again in a study by Faber et al. (2018) who detected phases of mind wandering based on fixations, saccades, blinks, and pupil size. They mention that these content-independent features work best for 12-s windows. Bixler and D'Mello (2016) compared the same features in a reading task with more task and content-specific features, such as repeated fixations on words. However, the general features performed better which allows for the conclusion that the general task-independent features could reach a good performance in other mind wandering and attention contexts as well. Several studies concentrated on gaining a further understanding on how fixations (Foulsham et al., 2013; Frank et al., 2015), saccades (Li et al., 2016), and eye blinks (Oh et al., 2012) are influenced by mental states. Features that were often extracted for the feature sets in the respective time interval include the number of fixations, saccades,

and blinks, as well as their average length, standard deviation, median, minimum, and maximum of the length, as well as angles between saccades and the ratio of fixations and saccades. Additionally, mean, standard deviation, median, minimum, and maximum were also calculated for the pupil diameter. However, Bixler and D'Mello (2016) note that the pupil diameter is very sensitive to luminance changes in the surroundings and requires a very careful and controlled setup. Nonetheless, the connection between mental states and the pupil diameter is also assessed in the studies by Franklin et al. (2013), Pfleging et al. (2016), Unsworth and Robison (2016), and Toker and Conati (2017). Mills et al. (2016) extended the mind wandering experiments to free viewing of films and found the same results for content-independent features compared to content-dependent features. The fixation and saccade features were also used in Hutt et al. (2017) who classified mind wandering during lecture viewing using a Bayes Net. In the mentioned studies by Faber et al. (2018) and Bixler and D'Mello (2016) many different algorithms were compared to find the best performance for the feature sets. For Faber et al. (2018) the highest performance was achieved with a Logistic Regression and for Bixler and D'Mello (2016) the best results were achieved by a Bayes Net and a Naïve Bayes algorithm.

A different feature set was tested by Xuelin Huang et al. (2019) who wanted to detect internal thought from eye vergence behavior features in three different tasks (math, watching a lecture video, and a daily activity like reading or browsing the internet). They used information from two different measures: pair-based vergence features and fixation-based vergence features. Their vergence feature set was compared to a feature set containing the previously mentioned features and the performance reached a similar level or even better results. If the features were combined, the best results were achieved. A comparison of several classification algorithms showed that a random forest yields the best results. It was suggested in Puig et al. (2013) that distinguishable eye vergence features are mainly related to covert visual attention tasks. In the literature, eye vergence features were found to be related to covert visual attention (Puig et al., 2013), imagination (Laeng and Sulutvedt, 2014) and internally and externally directed cognition (Benedek et al., 2017; Annerer-Walcher et al., 2020). Hence, eye vergence features are interesting features for the classification of attentional states.

Two of the data sets that are analyzed in this work focus on the classification of internal and external attention. Internally directed attention refers to attention that is independent of stimuli from the surroundings such as memory recall or mental arithmetic. Externally directed attention instead means focusing on sensory input, for example, visual search tasks or auditory attention to one of many speakers (Chun et al., 2011). Several studies found differences in eye behavior between internally and externally directed attention, especially for various features of pupil diameter, eye vergence, blinks, saccades, microsaccades, and fixations (e.g., Salvi et al., 2015; Unsworth and Robison, 2016; Benedek et al., 2017; Annerer-Walcher et al., 2020). Some features were more consistently associated with internally and externally directed cognition than others. It is hypothesized that two mechanisms mainly lead to the differences in eye behavior between internally and externally directed attention: decoupling of eye behavior from external stimuli (Smallwood and Schooler, 2006) and coupling of eye behavior to internal representations and processes (e.g., luminance and distance, Laeng and Sulutvedt, 2014). A detailed review of the general occulometric features that were mentioned before during internal and external attention was described in Annerer-Walcher et al. (2021). In Vortmann et al. (2019b), the authors implemented a real-time system that classifies internal and external attention based on multimodal EEG and eye tracking data. For the eye tracking data they used the previously described standard features (fixations, saccades, blinks, and pupil diameter), and classified short sequences of 3 s using a Linear Discriminant Analysis (LDA). This real-time classifier was later implemented in an attention-aware smart home system to improve the usability (Vortmann and Putze, 2020).

1.2. Related Work on Deep Learning for Eye Tracking

In more recent advances, deep learning approaches are used to improve different areas of eye tracking. Most of these studies do not focus on differentiating mental states from the data but rather improving the gaze estimation itself, unsupervised feature extractions, or predictions about the demographics of the participants. The use cases for the applications are many-fold, such as websites (Yin et al., 2018) or Augmented and Virtual Reality (Lemley et al., 2018).

As mentioned in the previous related work, the feature engineering for eye tracking classification remains a main research area. In Lohr et al. (2020), the authors explore using a metric learning approach to extract eye gaze features. They trained a set of three multilayer perceptrons to find fixations, saccades, and post-saccadic oscillations and reached benchmark performance for the detection. However, Bautista and Naval (2020) argue that extracting features based on fixations and saccades does not represent the richness of information available in eye tracking data. They suggest using deep unsupervised learning instead of feature engineering. Two autoencoders (AE) are trained on position and velocity information to extract macro-scale and micro-scale information and fitted the representations using a linear classifier. Their classification accuracy to discriminate gender and age groups reaches competitive levels compared to supervised feature extraction methods. Zhang and Le Meur (2018), instead, classified scanpaths using a one-dimensional CNN to predict the age of the participant.

Overall, using the scanpaths in the classification process instead of extracted statistical features can be observed in several recent studies. Assens et al. (2018) and Bao and Chen (2020) predict visual scanpaths using GANs and a deep convolutional saccadic model. In Fuhl et al. (2019), the scanpaths are represented by emojis in the first step. These representations were learned by a generative adversarial network (GAN). In a second step, the emojis are classified using a Convolutional Neural Network (CNN) to predict the stimulus. The authors argue that by adding the intermediate step of the emoji representation, they

increase the classification accuracy compared to classification simply based on scanpaths.

Sims and Conati (2020) used a combination of a Recurrent Neural Network (RNN) and a CNN to detect user confusion from eye tracking data. They argue that the parallel use of the neural nets allows keeping temporal information (using the RNN) and visuo-spatial information (using the CNN) and that their approach outperforms state-of-the-art classifiers. They used a 1-layer Gated Recurrent Unit (GRU) for the sequential eye tracking data and supplied the CNNs with scanpath images.

Another approach without explicit feature extraction was implemented by Zhang et al. (2019). They used a Deep Neural Network that was made up of several Long-Short-Term-Memories (LSTMs) to accurately detect Fetal Alcohol Spectrum Disorder in young children based on their natural viewing behavior.

Moving away from designated eye tracking devices, several studies have explored using other cameras for gaze detection. Different deep learning strategies have been applied in these studies to increase the tracking and classification accuracies of such systems. For example, Meng and Zhao (2017) used webcams and proposed to use five eye feature points for the tracking instead of only the iris center. These five points are detected using a CNN and afterward, another CNN is used to recognize different eye movement patterns. The iTracker by Krafka et al. (2016) is a CNN trained on a large-scale eye tracking dataset to predict gaze points without calibration based on the camera of a mobile device. It reaches state-of-the-art accuracy. CNN-based feature extraction for eye tracking using mobile devices was also assessed in Brousseau et al. (2020), where the authors suggest the combination of the camera with a 3D infrared model.

As mentioned before, Wang and Oates (2015) proposed to encode time series data as images and classify these images using CNNs. The resulting images could be a well-suited alternative to classical feature engineering for eye tracking, scanpaths, or raw data. The authors suggest two different approaches: Gramian Angular Fields and Markov Transition Fields. The two approaches are described in more detail in section 2.2.2. In their paper, they tested these two approaches as well as their combination on the twelve standard benchmark time-series datasets of language data and vital signs used in Oates et al. (2012) and compared them to state-of-the-art classifiers. The analysis showed that the new approaches reach similar results. Since then, their suggested methods have been applied in several other studies. In Thanaraj et al. (2020), the authors used the GAF successfully to classify EEG data for epilepsy diagnosis and in Bragin and Spitsyn (2019) GAF was used for motion imagery classification from EEG. We are not aware of eye tracking datasets that have been analyzed with MTF or GAF images.

2. METHODS

Pursuing the goal of a general assessment of the usability of the imaging time-series approach for eye tracking classification of attentional states, we decided to compare multiple classifiers on multiple data sets for their classification results. The datasets cover different aspects of attention and were either recorded for screen-based tasks or in Augmented Reality. Especially Augmented Reality devices with head-mounted displays offer a good opportunity to include an eye tracker in the headset and add an explicit or implicit option for user interaction. The latest generations of Augmented Reality devices even have built-in eye tracking. Available relevant work was used as a guideline to decide on the classifiers to compare. The general occulometric features that were mentioned in section 1.1 in combination with different classifiers that we found in earlier studies will be called "Statistical Summary Approaches" (see section 2.2.1). These 1D statistical summary approaches as classification algorithms will be compared with each other as well as with two different neural nets that were trained on a feature set that was generated by the Imaging Time Series approach from Wang and Oates (2015) (see section 2.2.2). Further, we evaluate different settings for the ITS approach as well as person- and task-dependence.

2.1. Data Sets

The three chosen data sets are different with regard to evoked attentional focus, mode of task presentation, tasks, number of recorded participants, and total number of trials and trial lengths. They were all recorded specifically targeting a binary classification between two states of attention. Two of the data sets were recorded during experiments that were controlled for internally and externally directed attention—two modes of attention that are usually alternated unconsciously in everyday life. The third data set contains trials of only externally directed visual attention. This visual attention is either directed toward real objects or virtual objects that are displayed by an Augmented Reality device. All three experimental tasks and setups will briefly be described in the following. All experiments were approved by their local ethics committees. Please refer to the original articles for a more detailed description. An overview of the data sets can be found in **Table 1**.

2.1.1. Switch-Task

The original research article of the switch-task data set was published in Annerer-Walcher et al. (2021). It was recorded as a cooperation of the University of Graz, Austria, and the University of Bremen, Germany. During the experiment, the participants were presented with 6 different types of tasks on a computer screen (see **Figure 1A** for task types). Each task was either numerical, verbal, or visuo-spatial and required either internally or externally directed attention. Participants were advised to keep their eyes open and focused on the screen, independent of the task. A task description was displayed before each trial. After a button press, a drift correction was performed while the participants focused on a fixation cross. For external attentional focus trials, it was necessary to attend the visual input on the screen and count the number of times the task could be answered with "yes." The shown stimulus always consisted of the elements necessary for all three external tasks and did not depend on the current task type (see **Figure 1B**). The trials lasted 10–14 s each and consisted of 8–11 stimulus screens of the same category. The trial length and type were chosen randomly. The stimulus screen (800 ms) was alternated with a masked screen (400 ms)

TABLE 1 | Overview of the three data sets including information about the tasks and scope.

Data set	Attention	Task presentation	Participants	Total trials	Trial length (s)
Switch	Internal/external	Screen-based	172	Approx. 15,000	10
Align	Internal/external	Augmented Reality	14	Approx. 900	15
Pairs	Real/virtual	Augmented Reality	13	Approx. 400	20

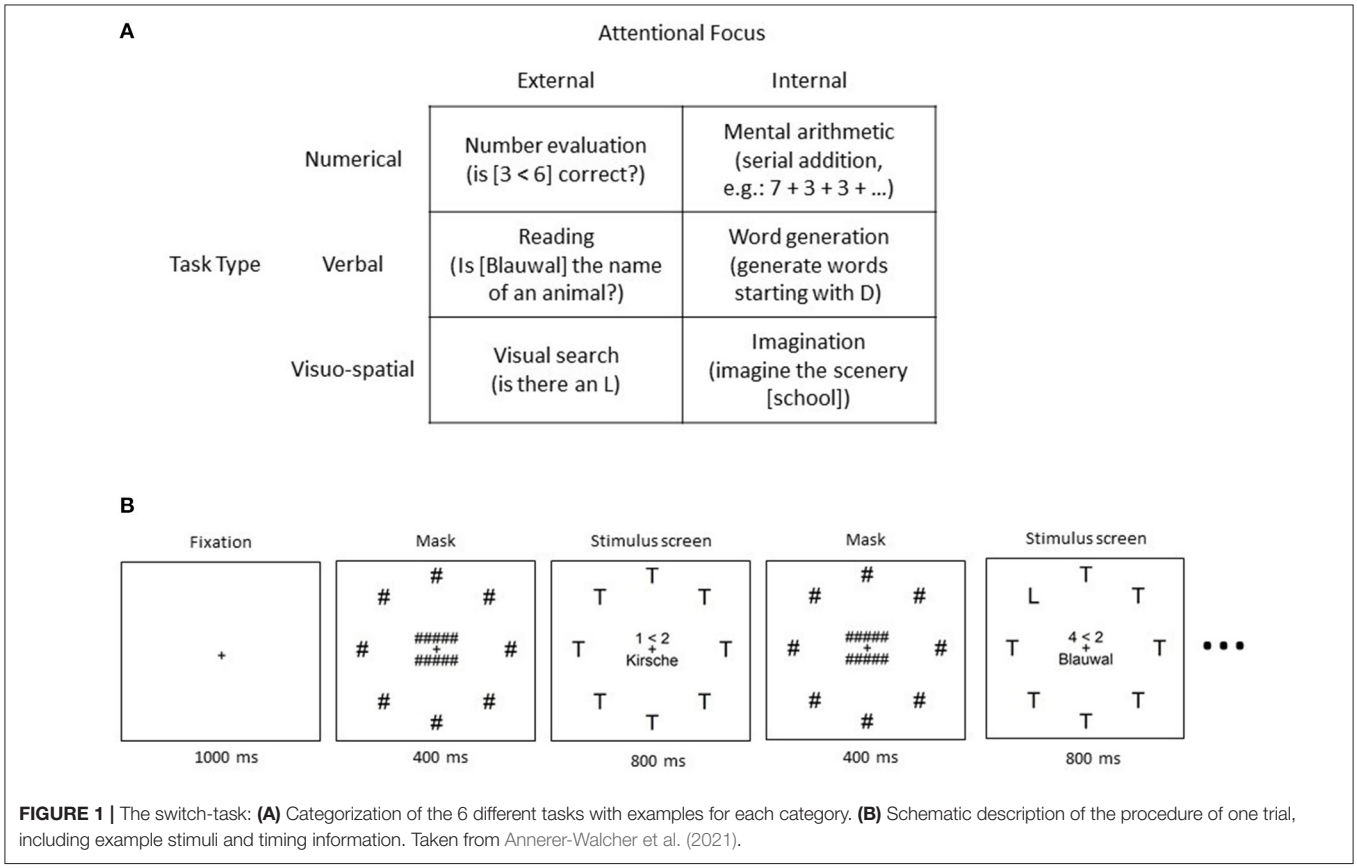


FIGURE 1 | The switch-task: **(A)** Categorization of the 6 different tasks with examples for each category. **(B)** Schematic description of the procedure of one trial, including example stimuli and timing information. Taken from Annerer-Walcher et al. (2021).

between the single tasks. For example, for an external numerical trial, the task was to count how many times the shown number comparison was correct (i.e., $9 < 7$). By always displaying a very similar visual stimulus, the differences between trials were minimized and restricted to the explicit task. Accordingly, the same presentation of visual stimulus screens was chosen for internal tasks even though their content was irrelevant for the tasks. An exemplary internal task was to generate as many words as possible starting with the letter D, without saying them out loud. Performance checks were randomly presented in 1/4 of the trials. A full data set of one participant consisted of two experiment blocks with 8 trials of each task in a randomized order (96 trials in total). Incomplete data sets were also included in our analysis.

For the binocular eye data recording an SMI RED250mobile system (SensoMotoric Instruments, Germany) with a temporal resolution of 250 Hz, spatial resolution of 0.03°, and gaze position accuracy of 0.4° visual angle was used. The participants' heads were stabilized using a chin rest.

2.1.2. Alignment-Task

In Vortmann et al. (2019a), the alignment-task of the second data set was described. In this study, internally and externally directed attention was evoked during an Augmented Reality scenario. The task of the participants was to visually align a virtual ball (red) and a virtual tube (green) that can be seen in **Figure 2A**. During the trials with externally directed attention, the ball kept moving in slow steady motions with direction changes every 5 s within a small distance from the center of the tube to keep the participant focused for 20 s. The tube was in a fixed position while the ball moved on a plane that was parallel to the surface of the tube but closer to the participant than the tube. The alignment was achieved by movement of the upper body and head. For the trials of internally directed attention, the participants learned to imagine the movement pattern of the ball based on a series of numbers. In a tutorial, the ball and/or a number pad were displayed in front of the tube (see schematic representation in **Figure 2B**). In the real internal trials, this number pad and ball had to be imagined by the participant. Before such a trial, a

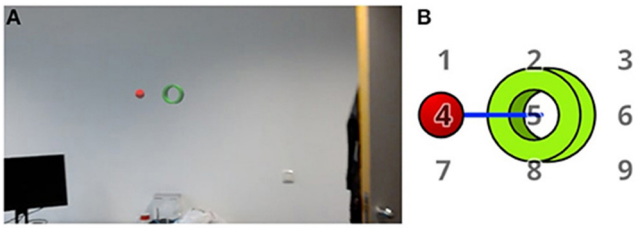


FIGURE 2 | The alignment-task: **(A)** Example scene from the HoloLens, showing green tube and unaligned red ball in front of a white wall. **(B)** Schematic display of the tube and the ball together with the imagined number pad for internal trials. The blue line indicates a ball movement from 4 to 5. Taken from Vortmann et al. (2019a).

sequence of 3 numbers between 1 and 9 was played as audio (i.e., 1-6-8). This sequence described the motion pattern of the imagined ball (i.e., upper left–middle right, lower middle). The participant's task was to imagine the movement and always slowly adjust their current position to keep the ball and tube aligned. They were advised to take approximately 5 s to imagine the movement of the ball from one number to the next number, resulting in a total trial time for internal trials of 15 s. Taken together, the task was always to keep the visual or imagined ball "inside" the tube by adjusting one's position. This task design was chosen to have two identical conditions regarding movement and visual input type while differing in the state of attention.

Participants performed 36 internal and 36 external alternating trials in total, split up into 3 blocks with breaks in between. The holograms and sounds were displayed using a Microsoft HoloLens 1. A binocular PupilLabs eye tracker with a sampling rate of 120 Hz was attached to the screen of the HoloLens to record the eye gaze. The average eye tracker accuracy is not available for this dataset.

2.1.3. Pairs-Task

The third data set was recorded during the performance of a pairs-task that was described in Vortmann et al. (2021). For this experiment, the participants had to play the children's game "pairs" with two different conditions in Augmented Reality. During the game, the participants have to memorize the positions of several cards. Each picture is present twice. These two cards are a pair and have to be identified as such while the cards are turned over to their neutral side with no pictures on them. In the first condition, the cards are real wooden cards while some of the surrounding elements are augmented content. In the second condition, the same cards with similar symbols are virtually added to the scene (see **Figure 3**). During the "memory"-phase, the participants see a deck of cards with the picture side up for 20 s and have time to memorize as many of the pairs as possible (varying deck sizes for different difficulties). Afterward, in the "remember"-phase, the participants can choose the pairs that they remembered. For the classification task, only the "memory"-phase will be regarded. During these 20 s, it can be assumed that the participants exclusively pay attention to the real or virtual cards, depending on the condition. Because the task is exactly the same in both conditions, the same viewing strategy would be assumed. With this data set, the goal is to see whether it is possible to classify attention on real vs. on virtual objects in Augmented Reality settings based on eye tracking data.

The same setup of the HoloLens 1 and the PupilLabs eye tracker as in the alignment task was used in this setup. The participants performed 20 trials of each condition. Trials with technical problems were excluded from the analysis. The average eye tracking accuracy after the calibration was 2.49 ± 0.51 degrees and on average 0.4 trials were excluded.

2.2. Classification Algorithms

To classify the different trial conditions in the presented data sets, different features, feature sets, and classification algorithms can be combined to optimize the classification performance. The goal of this study is to improve attentional state classification accuracy based on eye tracking data by following a new Imaging Time Series approach for the feature extraction. We will first describe which features were extracted for the statistical summary approach that was inspired by state-of-the-art related studies and will be used as a benchmark to compare the new approach to. This 1D feature set will be used to train several different classification algorithms. The ITS approach will contain a feature matrix of several generated images that will be used to train two different convolutional neural networks, which we will describe in section 2.2.2. No further preprocessing was applied to any of the datasets and no trials were excluded, other than already mentioned in section 2.1.3.

2.2.1. Statistical Summary Approaches

The general task-independent eye tracking features that are usually extracted were described in section 1.1. Which features can be extracted from the data sets is restricted by the format of the variables and values that were recorded by the eye trackers during the experiments. For some of the vergence features suggested by Xuelin Huang et al. (2019) information about the distance between the eyes and the distance between the focused object and the eyes is necessary. However, these are not given for all our data sets and thus we decided to combine the statistical summary feature set from fixations, saccades, blinks, remaining vergence features, and pupillometric data. For the extraction of these features, the data sequences of X and Y coordinates were evaluated for fixations, saccades, and blinks using the PyGaze Toolbox (Dalmaijer et al., 2014). The threshold value for the blink detection algorithm was 50 ms. Fixations were detected following the dispersion threshold identification algorithm (I-DT) by Salvucci and Goldberg (2000) (Implementation on github[1]). The dispersion threshold was set to 1 degree, as suggested by Blignaut (2009). The remaining vergence features were extracted as described in Xuelin Huang et al. (2019) and the minimal bounding circles were calculated with the python script from the nayuki-project[2]. As a feature, we either used the total value of the calculated variable, if possible (i.e., number of saccades),

[1]https://github.com/ecekt/eyegaze (assessed December, 2020).
[2]https://www.nayuki.io/page/smallest-enclosing-circle (assessed December, 2020).

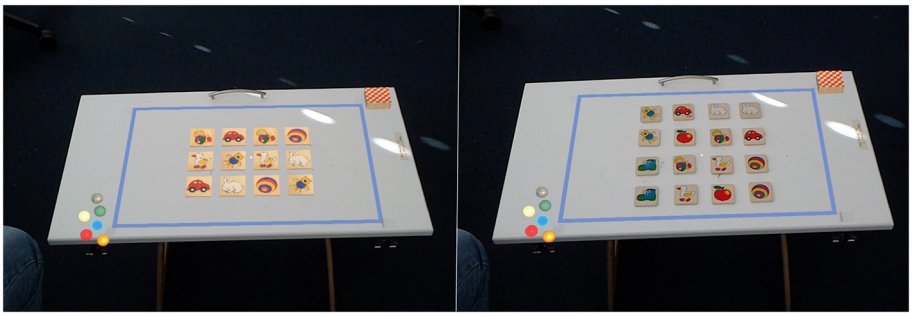


FIGURE 3 | The pairs-task: screenshots from the HoloLens showing the setup of the game. Virtual marbles and a deck of cards are always visible. On the **left** image the pairs cards are virtual, on the **right** image the cards are real. Taken from Vortmann et al. (2021).

or calculated statistical measures to describe the variable during the trial (i.e., mean, standard deviation, median, maximum, minimum, range, kurtosis, and skewness of the distribution of saccade lengths). For a complete list of all 76 features see the **Appendix**.

After feature extraction, all features are normalized using a z-score normalization. Features are ranked using an ANOVA estimator and a non-parametric mutual information estimator. These feature selection approaches were implemented using the scikit-learn toolbox by Pedregosa et al. (2011). As a hyperparameter optimization, we used the 10, 20, 30, 40, 50, 60, and 70 highest ranked features of both estimators.

The classification **algorithms** were also implemented using the default implementations from scikit-learn. We implemented the pipeline with the following algorithms:

- Naïve Bayes (NB)
- Logistic Regression (LogReg)
- Random Forest (RF)
- k-Nearest-Neighbor (knn)
- Linear Support Vector Machine (linSVM)
- Multi Layer Perceptron (MLP)
- and AdaBoost

The best feature set was chosen for each classifier individually by computing the average classification accuracy of all folds during five-fold cross-validation. The whole pipeline can be seen in **Figure 8** in the counter-clockwise path. This approach is used to gain optimal performance out of the classical approach, not considering any side-effects that could be caused by multiple testing of many classifier and feature set combinations (as they can only be beneficial for the classifiers and you are mainly interested in an upper bound).

2.2.2. Imaging Time Series

For the ITS approach, the continuous X and Y coordinate variables were transformed into images and classified using a neural net. In a preliminary step, phases during which blinks were detected were filtered from the data, because no information about the X and Y coordinates is available. A detailed description of the methods can be found in Wang and Oates (2015).

We decided to generate the images separately for the right and the left eye with one image representing the X coordinate and one image representing the Y coordinate recorded by the eye tracker. This way, we stay closest to visualizing the raw data and give the neural net the additional possibility to detect and learn from the differences and similarities between the eyes (following the idea of using vergence features). The first algorithm used for the transformation is the **Markov Transition Field** (MTF) which generates a matrix using transition probabilities. Based on the magnitude of the values, the data sequence S is split into Q quantiles. Each data point x_i is assigned to a quantile and a $Q \times Q$ weighed adjacency matrix W is constructed by counting the transitions from sample to sample between quantiles through a first-order Markov chain along the time axis. This Markov transition matrix W is then normalized and spread out among the magnitude axis considering the temporal positions, resulting in the MTF M. The main diagonal M_{ii} shows the self-transition probability at each time step (see **Figure 4**).

Additionally, we will work with two different versions of the Gramian Angular Field transformation algorithm. The first is called **Gramian Angular Summation Field** (GASF) and the second is called **Gramian Angular Difference Field** (GADF). For both methods, the data sequence X is rescaled to $[-1, 1]$ and then represented in polar coordinates by encoding the data values x as the angular cosine and the according timestamp as the radius. Thus, the data sequence is transferred from the Cartesian coordinate system into the polar coordinate system which has the advantage that for all points we preserve the absolute temporal relation. In the final step, we calculate the trigonometric sum (using cosine for the GASF) or the trigonometric difference (using sine for the GADF) pairwise between the points to identify the temporal correlation within time intervals. Accordingly, the Gramian matrix G has a size of $n \times n$ with n = length of raw time series. Each cell g_{ij} of G represents the trigonometric difference/sum of the points x_i and x_j with respect to the time interval. On the main diagonal, each cell g_{ii} contains the original value/angular information and could be used to reconstruct the original time series X. The steps of this algorithm are visualized in **Figure 5**, where Φ represents the time series in polar coordinates.

To reduce the size of the generated images, Piecewise Aggregation Approximation (PAA) can be applied for blurring

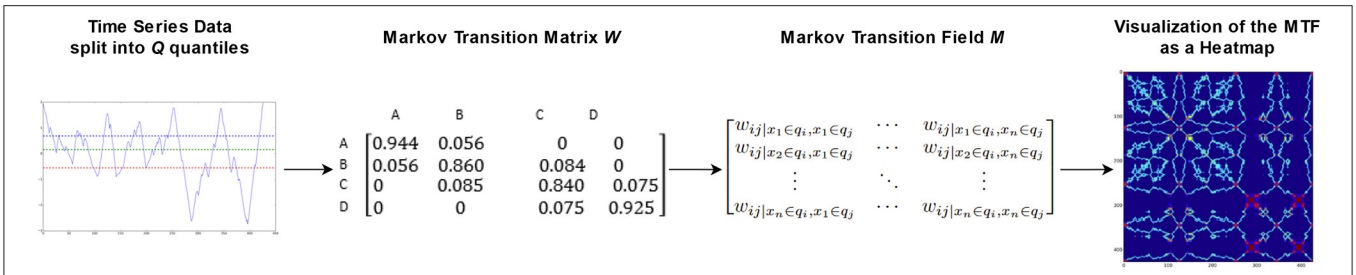


FIGURE 4 | Flowchart representing the main steps of the MTF algorithm. Adapted from Wang and Oates (2015).

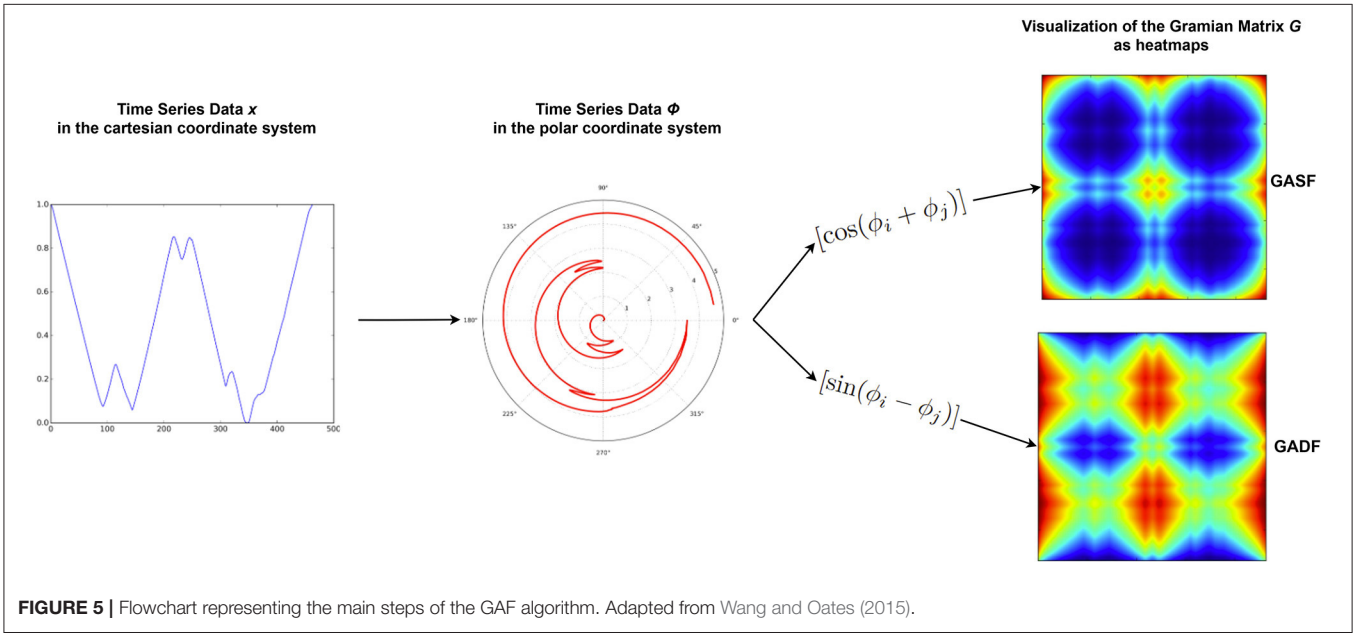


FIGURE 5 | Flowchart representing the main steps of the GAF algorithm. Adapted from Wang and Oates (2015).

(Keogh and Pazzani, 2000). The effect of blurring will be discussed in section 3.1.1.

The transformations of the data sequences into the MTF, GASF, and GADF images were implemented using the pyts-toolbox for python (Faouzi and Janati, 2020). The image size was set to 48*x*48 pixels and all pixel values were normalized between $[-1, 1]$ for individual images. Afterward, all generated images (3 transformations × 2 eyes × X/Y-coordinates = 12 images) were combined into an image matrix of size 3*x*4. This image generation process was applied to valid (non-blink) data of single trials per condition. An example of the images representing the feature matrix for an external trial of the switch-task data set can be seen in **Figure 6A**.

For the classification of the resulting images, we chose two CNNs with different complexities. The first CNN will be called **SimpleNet** and was implemented following the suggestions of Yang et al. (2020). It is made up of two convolutional layers with a kernel size of 5*x*5, two Max Pooling layers with a window size of 2*x*2 pixels, and two fully connected layers as well as the output layer. The number of units of the output layer is identical to the number of possible classification labels (in our cases: 2). Additionally, a dropout layer was included that temporarily freezes learned weights to avoid overfitting (see **Figure 7** for a schematic representation of the SimpleNet).

The second CNN is the **AlexNet** (Krizhevsky et al., 2017) that won the ImageNet Large Scale Visual Recognition Competition in 2012 (trained from scratch). It is more complex than the SimpleNet as it consists of 5 convolutional, 3 max-pooling, and 3 fully connected layers that are initialized with more channels/units. The learnable parameters in the AlexNet (57,081,730) are 41 times as many as in the SimpleNet (1,364,942). As in the statistical summary approach, the CNNs were trained in a five-fold cross-validation.

2.3. Analysis

For the classification, all trials of one data set were cut to the same length to avoid that the classifiers learn length-related information instead of attention-related information. That means, all trials of the switch data set were cut off after 10 s (equal task contribution was given) and the alignment-task data was shortened to 15-s windows for both conditions. The trials in the pairs data set were all equal in length and were thus kept at 20 s.

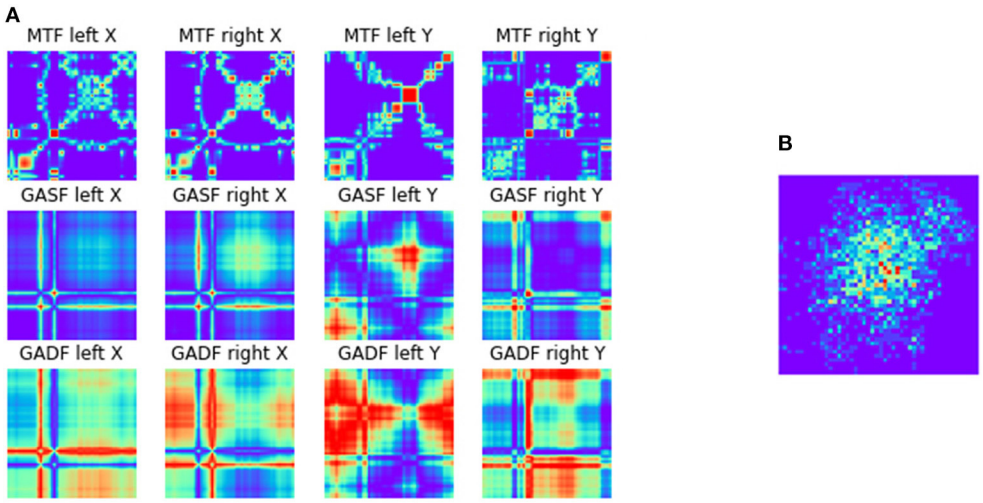


FIGURE 6 | (A) Exemplary feature matrix (3 × 4) made up of 12 images generated during the ITS approach as described in section 2.2.2. An external numeric trial from the switch-task is represented. Each row represents one of the transformation algorithms with one image for each eye/axis combination. **(B)** Heatmap of the gaze points representing the same trial.

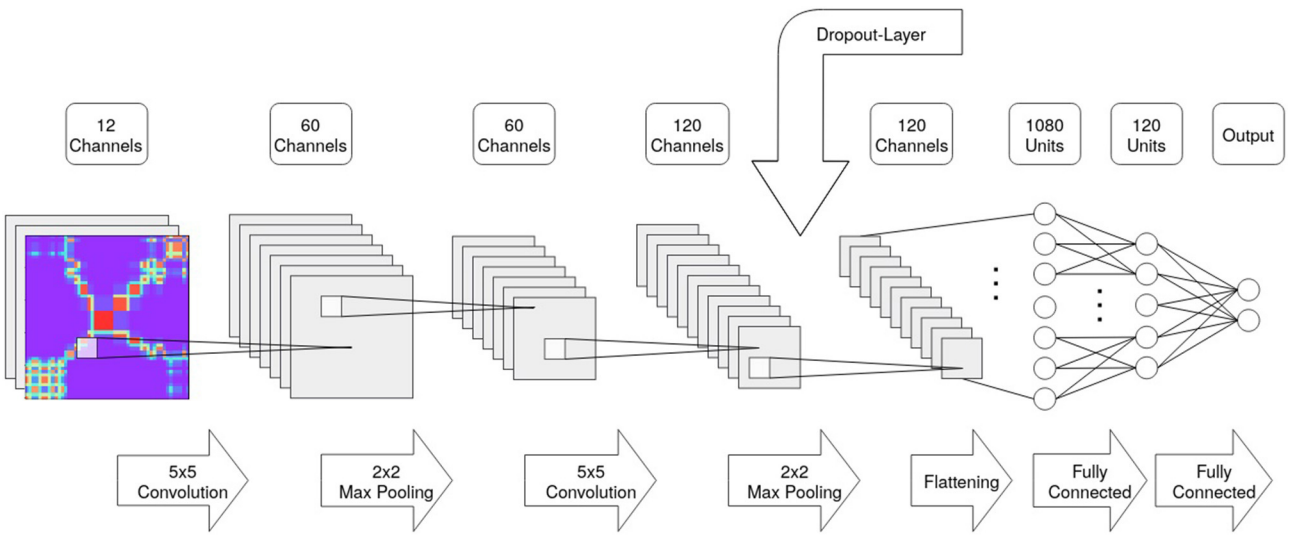


FIGURE 7 | Schematic description of the SimpleNet structure indicating layer size and number of channels/units.

The full pipeline—from the data sets to the comparison of the classifiers—can be seen in **Figure 8**. The counter-clockwise path shows the statistical summary approach and the clockwise path shows the ITS approach. As a performance metric, we chose to compare the resulting classification accuracies. This is possible because the attentional states were represented equally in the data sets. Accordingly, the chance level accuracy of guessing the correct attentional state for the binary classification tasks was 0.5.

For comparison of the classification accuracies, we want to determine whether one algorithm offers a statistically significant improvement over another approach. Therefore, we used a Wilcoxon-Signed Rank Test with a significance level of $\alpha = 0.05$. Paired data sets were assured by using reproducible training-test-splits across classifiers. Since we want to test whether one algorithm is not just different, but actually better (in this case returning lower values) than the other algorithm, we use the one-tailed version.

In the following, all comparisons will be presented in tables displaying the p-values of the one-tailed Wilcoxon-Signed-Rank Test. If a $p < 0.05$ is reported, that means that the classifier in that row performed significantly better than the classifier in that column. All values were rounded to 3 decimal places, thus, values of 1 and 0 are possible (0 meaning highly significant

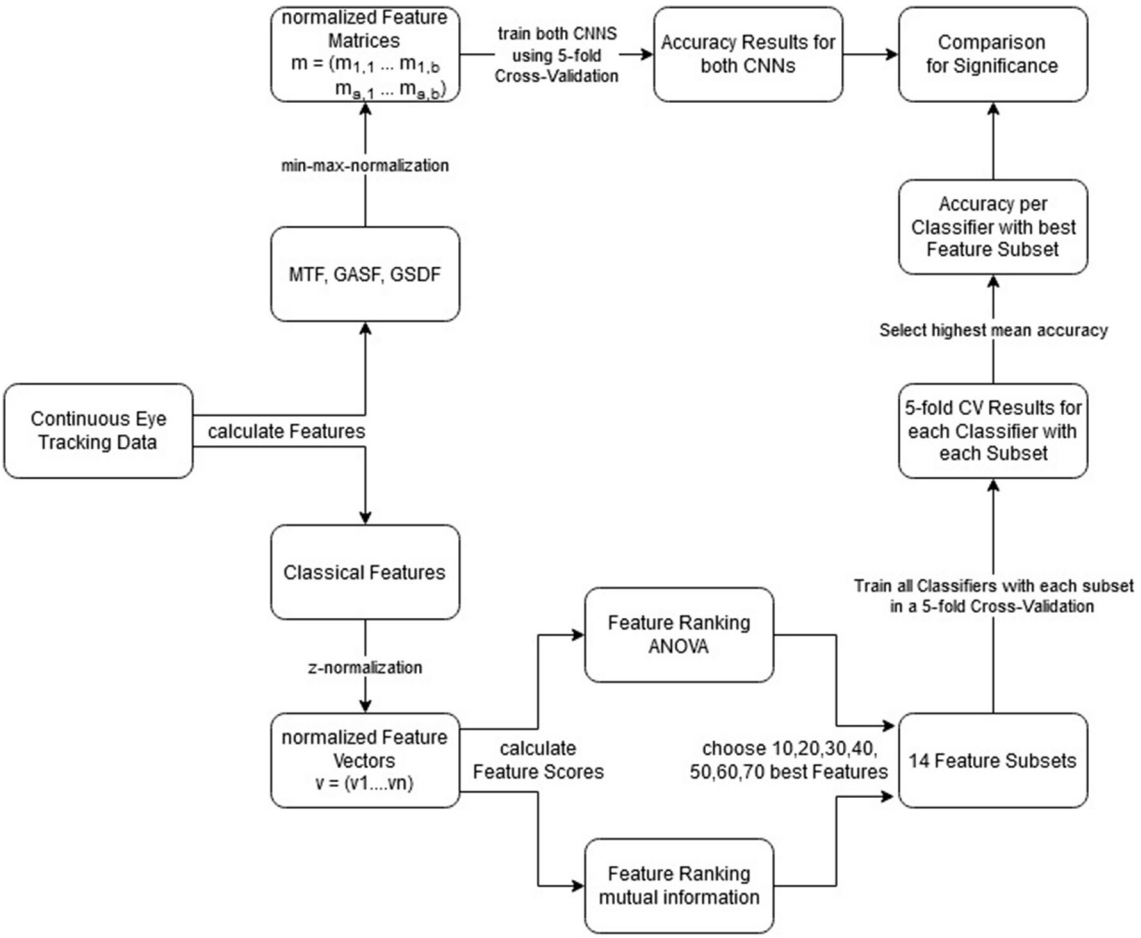


FIGURE 8 | Combined pipelines of the statistical summary and the ITS approach. Explaining step by step what was done to get from the continuous eye tracking data to the performed comparisons. Clockwise, ITS approach; Counter-Clockwise, statistical summary approach.

improvement). It follows that, if there is no $p < 0.05$ in one row, the classifier in that row did not perform significantly better than any other classifier. A "better" performance means a more accurate classification. Additionally, we report the mean classification accuracies for all classifiers in the tables.

For the training and testing splits, we followed three different strategies to answer three different research questions regarding the generalizability of the data. First, we train and test individually on data from the same participant (person-dependent, section 2.3.1). Afterwards, we test how well the data generalizes over participants (person-independent, section 2.3.2) and over tasks (task-generalizability, section 2.3.3).

2.3.1. Person-Dependent Classification

A person-dependent classifier is trained on data from one person and used to classify other data of the same person. For this approach, we took a participant's data from one dataset and performed a five-fold cross-validation with each of the suggested classification algorithms. For the statistical comparisons, the mean classification accuracy over the folds per participant was compared.

For reasons of computational time, only the SimpleNet was used with the ITS features during this analysis. The results are reported in section 3.2 and **Table 2**.

2.3.2. Person-Independent Classification

The person-independent version of the classifiers is trained on data that is independent of the participants whose data it is tested on. For this analysis, a combined data set over all participants per task is split and trained/tested using a group-five-fold cross-validation. That means the five-folds are chosen in a way that the data from one participant can never be in the training and in the testing data subset of that fold. The statistical comparisons are performed on the accuracy results of the individual folds.

The results are reported in section 3.3 and **Table 3**.

2.3.3. Task-Generalization

The switch-task data set contains an equal share of trials from 6 different tasks, 3 of which require internally directed attention and 3 of which require externally directed attention. As a final

TABLE 2 | Person-dependent: Average classification accuracies over all participants if the classifier was trained in a person-dependent manner; ***bold and italic***, highest average accuracy for this task; **bold**, *p*-value of the one-sided Wilcoxon Signed Rank Test above 0.05, thus no statistical difference between this and the best performing classifier.

	simpNet	knn	linSVM	RF	MLP	AdaBoost	NB	LogReg
Switch	***0.694***	0.609	0.619	0.58	0.571	0.559	0.612	0.604
Align	***0.707***	**0.667**	0.633	0.579	0.628	0.617	0.632	0.601
Pairs	***0.662***	0.589	**0.647**	0.585	0.614	0.524	**0.652**	0.582

TABLE 3 | Person-independent: Average classification accuracies over all folds of the group-five-fold cross-validation for the person-independent classifier; ***bold and italic***, highest accuracy for this task; **bold**, *p*-value of the one-sided Wilcoxon Signed Rank Test above 0.05, thus no statistical difference between this and the best performing classifier.

	simpNet	alexNet	knn	linSVM	RF	MLP	AdaBoost	NB	LogReg
Switch	***0.743***	0.73	0.642	0.685	0.674	0.688	0.69	0.619	0.689
Align	0.619	***0.705***	0.602	0.596	0.609	**0.641**	0.606	0.555	0.603
Pairs	0.52	0.5	**0.778**	**0.783**	**0.793**	**0.806**	**0.802**	0.715	***0.808***

TABLE 4 | Switch-task results, task-generalization: Average classification accuracies over all participants if the classifier was trained in using a LOOCV for each task in the switch dataset; ***bold and italic***, highest average accuracy; **bold**, *p*-value of the one-sided Wilcoxon Signed Rank Test above 0.05, thus no statistical difference between this and the best performing classifier.

	simpNet	alexNet	knn	linSVM	RF	MLP	AdaBoost	NB	LogReg
LOOCV	***0.783***	0.764	0.663	0.69	0.681	0.707	0.7	0.62	0.693

analysis, we wanted to test how the classifiers perform when they have to generalize over tasks. Analogously to the person-independent approach, we test the classifier on a task that it has not been trained on in a leave-one-out cross-validation (LOOCV). For example, we train the classifier using all trials, over all participants from the three external tasks and the numeric and verbal internal tasks but we test whether it correctly classifies all trials from the internal visuo-spatial task as internal. To do this, we chose a leave-one-task-out cross-validation. Again, the statistical analyses are performed on the accuracies of the folds.

The results can be seen in section 3.4 and **Table 4**.

3. RESULTS

Before the final comparison of all classifier implementations as described in section 2.3, we performed some preliminary tests to verify our approach and test the configurations regarding the optimal resolution of the images for the ITS approach.

3.1. Preliminary Tests

As suggested by Wang and Oates (2015), a blurring kernel can be used to decrease the resolution of the resulting images of the MTF, GASF, and GADF transformations. We were interested in how far a smaller image would lessen the classification accuracy because smaller images would lead to a reduced computation time (see section 3.1.1). Additionally, aiming at explainable AI, we had a look at the learned filters of the CNNs to assess whether the learned information is comparable to what is learned during image classification of real-world objects and whether we can understand what the CNN learns (see section 3.1.2). To test the hypothesis that the classifiers learn something about the differences between the conditions simply from different placements of the tasks in the visual field, we also trained our SimpleNet using heatmaps of the gaze coordinates and compared the results to the ITS approach (see section 3.1.3).

3.1.1. Image Resolution

To test the effect of the image resolution, we chose the same training and testing approach as described to the person-independent classifier (see section 2.3.2). We compared an image size of 12 × 12, 24 × 24, 36 × 36, and 48 × 48 pixels on the switch- and the alignment-task data sets as examples. Because the overall results of the pairs-data set were not significantly better than chance, we did not perform this comparison on this data set.

For both data sets, we find a better classification performance for a higher image resolution. For the switch-task data set, the classification accuracy improves significantly with a higher resolution up to a resolution of 36 × 36 pixels ($p = 0.0156$ compared to 24 × 24 pixels). Images with a resolution of 48 × 48 pixels lead to a higher mean accuracy with a lower variance, however, this improvement was not significant for our comparison.

For the alignment-task data set, the classification performance does not improve significantly for resolutions higher than 24 × 24 pixels. However, the mean accuracy still increases and the variance decreases with higher resolutions.

For the following analyses, we used an image resolution of 48 × 48 pixels because our computation time was of minor importance. However, if this approach is used in other

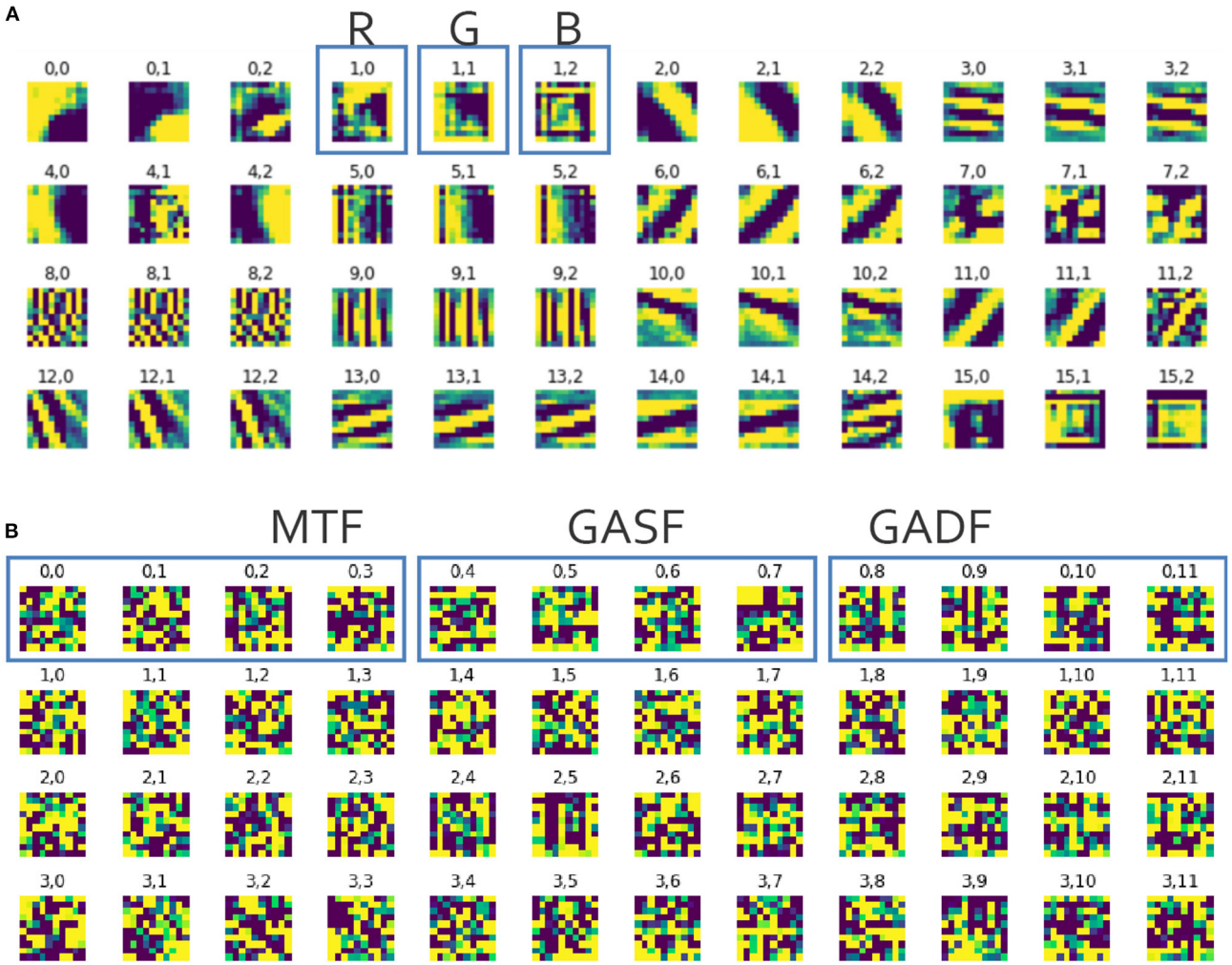


FIGURE 9 | Excerpt from the visualization of the filters learned by the first convolutional layer of **(A)** the AlexNet trained on an image data set of real objects (i.e., animals and fruit) where X.0-X.2 represent the RGB-channels, and **(B)** the SimpleNet trained on ITS features, where X.0-X.3 represent the MTF images, X.4-X.7 represent the GASF images and X.8-X.11 represent the GADF images.

studies, smaller image sizes can be chosen without significant performance loss.

3.1.2. Feature Analysis

The main reasoning behind using images that represent information from the raw data is that the Neural Net can abstract features that would not have been represented by an explicitly defined feature set. However, this is often argued to be a black box approach because it only tells us that there is a difference in the data but not what that difference is. Learning from clearly defined feature sets often allows for a detailed analysis on the importance of single features and thus, which features contain information about the differences between the conditions.

If a CNN is trained on images with real objects, the learned features often represent lines, edges, and other shapes (Krizhevsky et al., 2017). We visualized the features that were learned by the SimpleNet and found no such clear shapes or any other pattern that would explain what the CNN is learning from the ITS feature matrices (see **Figure 9**).

3.1.3. Heatmap Analysis

To shine some light on the question of whether the CNN abstracts pure spatial information from the ITS features, we generated heat maps for all the trials of all data sets and compared the achieved classification accuracies for the person-independent approach using the SimpleNet. An exemplary heatmap can be seen in **Figure 6B**. For the alignment- and the pairs-task data set, the classification performance was not significantly different from chance level (0.5). For the switch task data set the classification reached an average over all folds of 0.631 which suggests that there is some spatial information in the data set that allows for a differentiation between the internal and external condition. These results will be discussed further in section 4.

3.2. Person-Dependence

For the switch-task data set, the person-dependent classifiers were trained on approximately 70 trials in each fold. The highest mean classification accuracy of 69.4% across all participants was reached by the ITS-SimpleNet classifier. This result is significantly better than all statistical summary approaches. The second-best classification result was achieved by the linear Support Vector Machine (SVM) classifier with 61.9%.

The training subsets for the alignment task contained approximately 55 trials. Again, the highest classification accuracy was reached by the SimpleNet with 70.7% correctly classified trials on average. In this case, it was not significantly better than the best performing statistical summary classifier, which was the k-Nearest Neighbors approach with 66.7%. The SimpleNet is significantly better than all other tested classifiers.

In the pairs-task data set, the training subset for the person-dependent classifiers includes approximately 30 trials. As for the other two data sets, the highest classification accuracy is reported for the SimpleNet (66.2%) but with no significant improvement compared to the Naïve Bayes algorithm (65.2%) and the linear SVM (64.7%) (see **Table 2**).

Taken together, the SimpleNet reached the highest average accuracy for all three data sets if tested person-dependently with a significant improvement over all other statistical summary classifiers for the switch-task.

3.3. Person-Independence

Due to the combined data of the participants, the training subsets of the switch-task data set comprised approximately 12,000 trials for every fold in the person-independent approach. The classifiers that were trained on the ITS feature set performed significantly better than any of the statistical summary classifiers. The SimpleNet outperformed the AlexNet significantly with an accuracy of 74.3% compared to 73%. Of the statistical summary approaches, the linear SVM, the Random Forest, the Multi Layer Perceptron, the AdaBoost, and the Logistic Regression all classified approximately 68% of the trials correctly with no significant improvement over each other.

For the alignment data set, the combined trials result in training subsets of approximately 720 trials. The AlexNet had the highest classification accuracy of 70.5% on average over the folds. Only the Multi Layer Perceptron was not significantly worse with an accuracy of 64%.

The person-independent data set of the pairs-task resulted in approximately 320 training trials for each fold. The statistical summary approaches—except for the Naïve Bayes—reached accuracies of up to 80% with no significant statistical improvements over each other. The SimpleNet and the AlexNet only reached accuracies around 50% which is comparable to guessing (see **Table 3**).

The results show, that it is possible for all three data sets to generalize over all participants. However, which feature set captures the differences and similarities best is highly dependent on the attentional states that are to be classified.

3.4. Task-Generalizability

For the last analysis, the task independence of the features was tested by combining the switch-task trials of all participants and testing on only one of the six tasks. This resulted in approximately 12,500 trials in the training set. The best classification accuracy was achieved using the SimpleNet. It classified on average 78.3% of the trials correctly as internal or external attention even though it had never learned on trials from that task. This was significantly better than all the other classifiers. The second best classifier was the AlexNet with an accuracy of 76.4% which was significantly better than all statistical summary approaches. The best statistical summary approaches were the Multi Layer Perceptron, AdaBoost, and the Logistic Regression with up to 70.9% (see **Table 4**).

4. DISCUSSION

To optimize the accuracy of attentional state classification based on eye tracking data, different methods of feature extraction for various feature sets in combination with several classifiers have been tested in the past. In this work, we followed a new path by using an Imaging Time Series approach to visualize the raw eye tracking data and to classify the resulting images using convolutional neural networks. We compared the results with classical state-of-the-art approaches and found that our ITS approach outperforms the other classifiers. This difference can not be an advantage of deep learning in general, because the Multi Layer Perceptron that was trained on the statistical summary feature set was also significantly worse than the ITS approaches. However, a comparison between different image generation algorithms as features for the same deep learning classifier has yet to be assessed.

Even though the smallest amount of training data was used for person-dependently trained classifiers, the CNNs outperformed the general feature set classifiers in all three data sets. Interestingly, for the pairs data set, the CNNs that were trained person-independently on the ITS features did not achieve accuracies better than chance level, despite the bigger training data set. Since the classification was significantly better for the person-dependent classification, we assume that the ITS approach captures some characteristics of the eye gaze behavior that are different between the attention on real and virtual objects. However, the bad person-independent results suggest that the information that is captured in the ITS features is very individual between participants regarding viewing behavior. The statistical summary features and classifiers reached accuracies up to 80% for this task, thus, there are person-independent eye gaze feature differences during attention on real and virtual objects, these are just not learned in the ITS approach. Understanding this result requires further insight into the information that is encoded into the images and which filters were learned by the convolutional neural net. So far, the only conclusion we can draw from this is that the statistical features contain information that is missing in the ITS approach but would be important to classify attention on real and virtual objects in a person-independent manner. We excluded poorly randomized training

and testing data as the reason for the low classification accuracy by using the same splits across classification approaches. Also, the comparatively small amount of available data has a low probability of causing the low performance because the person-dependent classification for the pairs task was performed on even fewer data and reached a better performance.

For the two internal/external data sets the highest accuracy for the approach that generalizes over participants was again reached by one of the suggested new classification approaches over the statistical summary approaches. What can be noted is that in the switch data set, the SimpleNet performs significantly better than the AlexNet, while for the alignment data set it is the other way around. The results between the two CNNs are similar for the pairs- and the switch-task ($< 2\%$) but the accuracy for the SimpleNet used on the alignment data set is almost 9% worse than the AlexNet.

An interesting question that could be followed here is in how far the different complexities of the two models require different amounts of training data to reach similar results. The effect of more training data for CNNs was also discussed in Zhu et al. (2016) where they investigate the saturation threshold for the models. They conclude that while bigger data sets are almost always better, the real improvement happens when the representations of the data and the learning algorithms improve and are capable of profiting from larger data sets. While, a more complex model with more learnable parameters is more prone to overfitting if the amount of data is too small, it is also capable of capturing more complex structures. However, adding parameter complexity beyond the optimum reduces model quality. More training data is desirable because it reduces the variance in the model and displays more accurately which aspects of the data are general and which are the noise of specific trials. In our current analysis, we have not yet identified which characteristics of the two compared CNNs are responsible for the differences in the achieved classification accuracies. We assume, that the required complexity of the model is dependent on the attentional or in general mental states that are to be classified. This topic will need further investigation.

A very noticeable achievement is that the classification accuracies with the ITS approach for internal and external attention do not decrease for person-independent classification (74.3 and 70.2%) compared to person-dependent classification (69.4 and 70.7%) and for the pairs dataset it even increased (80.8% compared to 66.2%) when the Logistic Regression was chosen. For user applications that make real-time use of the classification results, a person-independent classifier eliminates the need for a long session of recordings just to train the classifier. This helps to develop real-time training-free use case scenarios where eye tracking data can be used to detect internally and externally directed attention in the user and if the attention is directed externally in Augmented reality settings, it can be classified whether the focus lies on real or virtual objects.

Another promising result is the high accuracy achieved for the task generalizability analysis. Using the ITS features together with the SimpleNet resulted in 78.3% correctly classified trials on average even though the classifier was not trained on data from that task. In Annerer-Walcher et al. (2021), the authors reported an accuracy of approximately 61% for their task transfer classification approach using an LSTM with the standard features. One difference is that they trained on two internal and two external tasks and tested on the remaining two. However, the classification accuracy reached by our approach is remarkably higher and we assume that not all of this difference can be explained by the different test/training split. We propose that the characteristics of the gaze behavior that are represented in the Imaging Time Series features are a good representation of what is shared over tasks during certain attentional states.

The trial lengths that were analyzed in this study (10–20 s) were adopted from the original studies for better comparability. To use the proposed methods in an online real-time system or for a temporally detailed offline classification, the approach should be adapted to either use smaller windows or sliding windows. While, smaller windows also reduce the available data for each decision, this is not the case for overlapping sliding windows. Appropriate window lengths or window overlaps for sliding windows highly depend on the context. While, some research questions might require a fine-grained analysis of attention switches (e.g., to study the exact steps of a single cognitive process), most applications would rather benefit from the detection of robust attention changes for longer periods (e.g., adapting a user interface to the attentional state, where too frequent changes would be more distracting than helpful).

Our study was the first to assess this classification approach for attentional states based on eye tracking. We were able to show an improvement in classification accuracy and are optimistic that further optimization can be achieved. A shortcoming of the presented analysis is that all the implemented classifiers were implemented in their default settings. Our goal was to use the same classifiers on all data sets and thus not optimize each classifier independently for each data set and classifier training variant. We are aware that the classification accuracy of the statistical summary approaches could be increased by performing further hyperparameter optimization additionally to the feature selection criteria. On the other hand, the CNNs that were used to classify the ITS features were also taken "out of the box" and were not optimized and designed specifically for this analysis. Typically, neural nets require a large amount of training data, which could be assessed in further experiments. We conclude that their results could be improved in the same dimensions that the statistical summary algorithms could be improved. Our goal was to show that this feature set is an interesting alternative that requires further attention because it might lead to better classifier performances.

A bigger challenge for the new approach is the interpretation of the model. While, the feature importance and differences can easily be analyzed for the statistical summary features, the parameters that are learned during the training of the CNNs with the images are harder to interpret. A pitfall of the ITS approach is its dependency on the gaze coordinates if these are the main difference for the learned conditions in the training data set. In the switch-task there seem to be differences between the conditions regarding the gaze heatmaps. A classifier should not learn that internally directed attention is present whenever the participants look to the left and externally directed

attention is present whenever the participants look to the right because it is not task and location independent. The statistical summary features do not fall for this information. In our case, the results of the person-independent ITS classification (74.3%) are significantly better than the results using a heatmap of the gaze coordinates (63.1%) which shows that the classifier learns significantly more from the Imaging Time Series than the "location."

All in all, the results of this first exploration of Imaging Time Series for eye tracking classification show that it is promising to further test and optimize in this direction, exploring other feature extraction and combination methods.

4.1. Future Work

In this work, the Imaging Time Series approach was tested on three different datasets. In the next step, other available eye tracking data sets of attentional states will be classified using this feature set. If possible, these data sets should contain other tasks and attentional states. The analyses will focus on understanding and optimizing the necessary complexity of the CNNs while keeping task- and person-independence in mind as a central goal.

After comparing the ITS approach to classical statistical gaze features, future comparisons will focus on other deep learning approaches that have been used on eye tracking data by related studies. In particular, we would be interested in a comparison of our suggested ITS approach with the approach from Sims and Conati (2020) where the CNNs were trained on the scanpaths and the temporal dimension was analyzed using GRUs.

Further, we want to investigate how well a combination of the statistical summary features and the ITS techniques mix. The statistical summary features contain a lot of information that is well-understood and can be explained by results from cognitive science research. However, with the statistical summary feature extraction and generation algorithms, a lot of information about the data is lost, especially with regard to the temporal dynamics within a trial. One idea would be to visualize some of the statistical summary features using Imaging Time Series. For example, the statistical summary features that describe the length of the saccades within a trial are often represented by statistical values that describe their distribution: Mean, standard deviation, minimum, and maximum. The saccade lengths are also a time series that could be transformed into an image with less information loss than the descriptive statistics. This could be an efficient combination of both approaches.

One last topic that was not addressed until now in this study is the window length of the classified data. With follow-up studies, we want to examine which effect the chosen time interval has on the classification accuracy. Precisely, shorter windows are desired if the accuracy loss is not significant because shorter trials would allow attentional state detection closer to real-time.

The overall goal will be an end-to-end system that can classify multiple aspects of the attentional state of a user without person-dependent training as fast and accurate as possible and use the information for adaptations of the interface or as implicit input.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are available by request. Requests to access these datasets should be directed to Lisa-Marie Vortmann, vortmann@uni-bremen.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee, University of Bremen. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The study was planned by L-MV, JK, and FP. The implementation was performed by JK. L-MV and JK analyzed and discussed the results. L-MV wrote the paper. SA-W, MB, and FP reviewed the paper. FP supervised the process. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Zentrale Forschungsförderung of the University of Bremen in the context of the project Attention-driven Interaction Systems in Augmented Reality. Open access was supported by the Open Access Initiative of the University of Bremen and the DFG. This work was supported by the Austrian Science Fund (FWF): P29801-B27 and P34043.

REFERENCES

Annerer-Walcher, S., Ceh, S., Putze, F., Kampen, M., Körner, C., and Benedek, M. (2021). How reliably do eye parameters indicate internal vs. external attentional focus? *Cognitive Science*, 45, 1–30. doi: 10.1111/cogs.12977

Annerer-Walcher, S., Körner, C., Beaty, R., and Benedek, M. (2020). Eye behavior predicts susceptibility to visual distraction during internally directed cognition. *Attent. Percept. Psychophys.* 82, 3432–3444. doi: 10.3758/s13414-020-02068-1

Assens, M., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2018). "Pathgan: visual scanpath prediction with generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (Munich). doi: 10.1007/978-3-030-11021-5_25

Bao, W., and Chen, Z. (2020). Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing* 404, 154–164. doi: 10.1016/j.neucom.2020.03.060

Bautista, L. G. C., and Naval, P. C. Jr. (2020). Gazemae: general representations of eye movements using a micro-macro autoencoder. *arXiv preprint arXiv:2009.02437*. doi: 10.1109/ICPR48806.2021.9412761

Benedek, M., Schickel, R. J., Jauk, E., Fink, A., and Neubauer, A. C. (2014). Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia* 56, 393–400. doi: 10.1016/j.neuropsychologia.2014.02.010

Benedek, M., Stoiser, R., Walcher, S., and Körner, C. (2017). Eye behavior associated with internally versus externally directed cognition. *Front. Psychol.* 8:1092. doi: 10.3389/fpsyg.2017.01092

Bixler, R., and D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Model. User Adapt. Interact.* 26, 33–68. doi: 10.1007/s11257-015-9167-1

Blignaut, P. (2009). Fixation identification: the optimum threshold for a dispersion algorithm. *Attent. Percept. Psychophys.* 71, 881–895. doi: 10.3758/APP.71.4.881

Bragin, A. D., and Spitsyn, V. G. (2019). "Electroencephalogram analysis based on gramian angular fieldtransformation," in *CEUR Workshop Proceedings* (Bryansk), 273–275. doi: 10.30987/graphicon-2019-2-273-275

Brousseau, B., Rose, J., and Eizenman, M. (2020). Hybrid eye-tracking on a smartphone with cnn feature extraction and an infrared 3d model. *Sensors* 20:543. doi: 10.3390/s20020543

Buscher, G., Cutrell, E., and Morris, M. R. (2009). "What do you see when you're surfing? Using eye tracking to predict salient regions of web pages," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA), 21–30. doi: 10.1145/1518701.1518705

Ceh, S., Annerer-Walcher, S., Körner, C., Rominger, C., Kober, S. E., Fink, A., et al. (2020). Neurophysiological indicators of internal attention: an EEG-eye-tracking co-registration study. *Brain Behav.* 10, 1–14. doi: 10.1002/brb3.1790

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol.* 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Conati, C., Aleven, V., and Mitrovic, A. (2013). Eye-tracking for student modelling in intelligent tutoring systems. *Design Recommend. Intell. Tutor. Syst.* 1, 227–236.

Dalmaijer, E. S., Mathôt, S., and Van der Stigchel, S. (2014). Pygaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behav. Res. Methods* 46, 913–921. doi: 10.3758/s13428-013-0422-2

Dehais, F., Dupres, A., Di Flumeri, G., Verdiere, K., Borghini, G., Babiloni, F., et al. (2018). "Monitoring pilot's cognitive fatigue with engagement features in simulated and actual flight conditions using an hybrid fNIRS-EEG passive BCI," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (Miyazaki: IEEE), 544–549. doi: 10.1109/SMC.2018.00102

Di Stasi, L. L., McCamy, M. B., Catena, A., Macknik, S. L., Canas, J. J., and Martinez-Conde, S. (2013). Microsaccade and drift dynamics reflect mental fatigue. *Eur. J. Neurosci.* 38, 2389–2398. doi: 10.1111/ejn.12248

Faber, M., Bixler, R., and D'Mello, S. K. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behav. Res. Methods* 50, 134–150. doi: 10.3758/s13428-017-0857-y

Faouzi, J., and Janati, H. (2020). pyts: a python package for time series classification. *J. Mach. Learn. Res.* 21, 1–6.

Foulsham, T., Farley, J., and Kingstone, A. (2013). Mind wandering in sentence reading: decoupling the link between mind and eye. *Can. J. Exp. Psychol.* 67:51. doi: 10.1037/a0030217

Frank, D. J., Nara, B., Zavagnin, M., Touron, D. R., and Kane, M. J. (2015). Validating older adults' reports of less mind-wandering: an examination of eye movements and dispositional influences. *Psychol. Aging* 30:266. doi: 10.1037/pag0000031

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., and Schooler, J. W. (2013). Window to the wandering mind: pupillometry of spontaneous thought while reading. doi: 10.1080/17470218.2013.858170

Fuhl, W., Bozkir, E., Hosp, B., Castner, N., Geisler, D., Santini, T. C., et al. (2019). "Encodji: encoding gaze data into emoji space for an amusing scanpath classification approach," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research Applications* (Denver, CO), 1–4. doi: 10.1145/3314111.3323074

Horng, W.-B., Chen, C.-Y., Chang, Y., and Fan, C.-H. (2004). "Driver fatigue detection based on eye tracking and dynamic template matching," in *IEEE International Conference on Networking, Sensing and Control, 2004* (Taipei: IEEE), 7–12. doi: 10.1109/ICNSC.2004.1297400

Huey, E. B. (1908). *The Psychology and Pedagogy of Reading*. New York, NY: The Macmillan Company.

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., and D'Mello, S. K. (2017). *Gaze-Based Detection of Mind Wandering During Lecture Viewing*. International Educational Data Mining Society.

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87:329. doi: 10.1037/0033-295X.87.4.329

Keogh, E. J., and Pazzani, M. J. (2000). "Scaling up dynamic time warping for datamining applications," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, MA), 285–289. doi: 10.1145/347090.347153

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., et al. (2016). "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2176–2184. doi: 10.1109/CVPR.2016.239

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Laeng, B., and Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychol. Sci.* 25, 188–197. doi: 10.1177/0956797613503556

Lemley, J., Kar, A., and Corcoran, P. (2018). "Eye tracking in augmented spaces: a deep learning approach," in *2018 IEEE Games, Entertainment, Media Conference (GEM)* (Galway: IEEE), 1–6. doi: 10.1109/GEM.2018.8516529

Li, J., Ngai, G., Leong, H. V., and Chan, S. C. (2016). "Your eye tells how well you comprehend," in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)* (Atlanta, GA: IEEE), 503–508. doi: 10.1109/COMPSAC.2016.220

Lohr, D., Griffith, H., Aziz, S., and Komogortsev, O. (2020). "A metric learning approach to eye movement biometrics," in *2020 IEEE International Joint Conference on Biometrics (IJCB)* (Houston, TX: IEEE), 1–7. doi: 10.1109/IJCB48548.2020.9304859

Meng, C., and Zhao, X. (2017). Webcam-based eye movement analysis using CNN. *IEEE Access* 5, 19581–19587. doi: 10.1109/ACCESS.2017.2754299

Miller, S. A. (2020). *Eye Tracking Systems and Method for Augmented or Virtual Reality*. US Patent 10825248.

Mills, C., Bixler, R., Wang, X., and D'Mello, S. K. (2016). *Automatic Gaze-Based Detection of Mind Wandering During Narrative Film Comprehension*. International Educational Data Mining Society.

Oates, T., Mackenzie, C. F., Stein, D. M., Stansbury, L. G., Dubose, J., Aarabi, B., et al. (2012). "Exploiting representational diversity for time series classification," in *2012 11th International Conference on Machine Learning and Applications* (Boca Raton, FL: IEEE), 538–544. doi: 10.1109/ICMLA.2012.186

Oh, J., Jeong, S.-Y., and Jeong, J. (2012). The timing and temporal patterns of eye blinking are dynamically modulated by attention. *Hum. Movement Sci.* 31, 1353–1365. doi: 10.1016/j.humov.2012.06.003

Paletta, L., Neuschmied, H., Schwarz, M., Lodron, G., Pszeida, M., Ladstätter, S., et al. (2014). "Smartphone eye tracking toolbox: accurate gaze recovery on mobile displays," in *Proceedings of the Symposium on Eye Tracking Research and Applications* (Safety Harbor, FL), 367–68. doi: 10.1145/2578153.2628813

Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. (2010). "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (Austin, TX), 141–144. doi: 10.1145/1743666.1743701

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pfleging, B., Fekety, D. K., Schmidt, A., and Kun, A. L. (2016). "A model relating pupil diameter to mental workload and lighting conditions," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA), 5776–5788. doi: 10.1145/2858036.2858117

Puig, M. S., Zapata, L. P., Aznar-Casanova, J. A., and Supér, H. (2013). A role of eye vergence in covert attention. *PLoS ONE* 8:e52955. doi: 10.1371/journal.pone.0052955

Salvi, C., Bricolo, E., Franconeri, S. L., Kounios, J., and Beeman, M. (2015). Sudden insight is associated with shutting out visual inputs. *Psychon. Bull. Rev.* 22, 1814–1819. doi: 10.3758/s13423-015-0845-0

Salvucci, D. D., and Goldberg, J. H. (2000). "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, FL), 71–78. doi: 10.1145/355017.355028

Sims, S. D., and Conati, C. (2020). "A neural architecture for detecting user confusion in eye-tracking data," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 15–23. doi: 10.1145/3382507.3418828

Smallwood, J., and Schooler, J. W. (2006). The restless mind. *Psychol. Bull.* 132:946. doi: 10.1037/0033-2909.132.6.946

Thanaraj, K. P., Parvathavarthini, B., Tanik, U. J., Rajinikanth, V., Kadry, S., and Kamalanand, K. (2020). Implementation of deep neural networks to classify EEG signals using gramian angular summation field for epilepsy diagnosis. *arXiv preprint arXiv:2003.04534*.

Toker, D., and Conati, C. (2017). "Leveraging pupil dilation measures for understanding users' cognitive load during visualization processing," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava), 267–270. doi: 10.1145/3099023.3099059

Unsworth, N., and Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cogn. Affect. Behav. Neurosci.* 16, 601–615. doi: 10.3758/s13415-016-0417-4

Vézard, L., Legrand, P., Chavent, M., Faita-Ainseba, F., and Trujillo, L. (2015). EEG classification for the detection of mental states. *Appl. Soft Comput.* 32, 113–131. doi: 10.1016/j.asoc.2015.03.028

Vortmann, L.-M., Kroll, F., and Putze, F. (2019a). EEG-based classification of internally-and externally-directed attention in an augmented reality paradigm. *Front. Hum. Neurosci.* 13:348. doi: 10.3389/fnhum.2019.00348

Vortmann, L.-M., and Putze, F. (2020). "Attention-aware brain computer interface to avoid distractions in augmented reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI EA '20* (New York, NY: Association for Computing Machinery), 1–8. doi: 10.1145/3334480.3382889

Vortmann, L.-M., Schult, M., Benedek, M., Walcher, S., and Putze, F. (2019b). "Real-time multimodal classification of internal and external attention," in *Adjunct of the 2019 International Conference on Multimodal Interaction, ICMI '19* (New York, NY: Association for Computing Machinery). doi: 10.1145/3351529.3360658

Vortmann, L.-M., Schwenke, L., and Putze, F. (2021). Real or virtual? Using brain activity patterns to differentiate attended targets during augmented reality scenarios. *arXiv [Preprint] arXiv:2101.05272.*

Wang, Z., and Oates, T. (2015). "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, TX).

Wedel, M., and Pieters, R. (2008). A review of eye-tracking research in marketing. *Rev. Market. Res.* 4, 123–147. doi: 10.1108/S1548-6435(2008)0000004009

Xuelin Huang, M., Li, J., Ngai, G., Leong, H. V., and Bulling, A. (2019). Moment-to-moment detection of internal thought from eye vergence behaviour. *arXiv preprint arXiv:1901.06572*. doi: 10.1145/3343031.3350573

Yang, C.-L., Chen, Z.-X., and Yang, C.-Y. (2020). Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images. *Sensors* 20:168. doi: 10.3390/s20010168

Yin, Y., Juan, C., Chakraborty, J., and McGuire, M. P. (2018). "Classification of eye tracking data using a convolutional neural network," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (IEEE), 530–535. doi: 10.1109/ICMLA.2018.00085

Zeng, H., Yang, C., Dai, G., Qin, F., Zhang, J., and Kong, W. (2018). EEG classification of driver mental states by deep learning. *Cogn. Neurodyn.* 12, 597–606. doi: 10.1007/s11571-018-9496-y

Zhang, A. T., and Le Meur, B. O. (2018). "How old do you look? Inferring your age from your gaze," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE), 2660–2664. doi: 10.1109/ICIP.2018.8451219

Zhang, C., Paolozza, A., Tseng, P. H., Reynolds, J. N., Munoz, D. P., and Itti, L. (2019). Detection of children/youth with fetal alcohol spectrum disorder through eye movement, psychometric, and neuroimaging data. *Front. Neurol.* 10:80. doi: 10.3389/fneur.2019.00080

Zhu, X., Vondrick, C., Fowlkes, C. C., and Ramanan, D. (2016). Do we need more training data? *Int. J. Comput. Vis.* 119, 76–92. doi: 10.1007/s11263-015-0812-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Complete list of features for the statistical summary feature set. sd, standard deviation; min, minimum; max, maximum.

- Distance between gaze points of both eyes mean
- Distance between gaze points of both eyes sd
- Angle between gaze points of both eyes mean
- Angle between gaze points of both eyes sd
- Distance between centroids of both eyes
- Angle between centroids of both eyes
- Distance between minimal bounding circles of both eyes
- Angle between minimal bounding circles of both eyes
- Normalized distance between minimal bounding circles of both eyes
- Minimal bounding circle radius left eye
- Minimal bounding circle radius right eye
- Fixation duration mean
- Fixation duration sd
- Fixation duration median
- Fixation duration min
- Fixation duration max
- Fixation duration range
- Fixation duration kurtosis
- Fixation duration skewness
- Fixation quantity
- Fixations total duration
- Saccade duration mean
- Saccade duration sd
- Saccade duration median
- Saccade duration min
- Saccade duration max
- Saccade duration range
- Saccade duration kurtosis
- Saccade duration skewness
- Saccade length mean
- Saccade length sd
- Saccade length median
- Saccade length min
- Saccade length max
- Saccade length range
- Saccade length kurtosis
- Saccade length skewness
- Saccade velocity mean
- Saccade velocity sd
- Saccade velocity median
- Saccade velocity min
- Saccade velocity max
- Saccade velocity range
- Saccade velocity kurtosis
- Saccade velocity skewness
- Saccade quantity
- Saccades total duration
- Angles between saccade and x-axis mean
- Angles between saccade and x-axis sd
- Angles between saccade and x-axis median
- Angles between saccade and x-axis min
- Angles between saccade and x-axis max
- Angles between saccade and x-axis range
- Angles between saccade and x-axis kurtosis
- Angles between saccade and x-axis skewness
- Angles between saccades mean
- Angles between saccades sd
- Angles between saccades median
- Angles between saccades min
- Angles between saccades max
- Angles between saccades range
- Angles between saccades kurtosis
- Angles between saccades skewness
- Fixation/saccade duration ratio
- Blink duration mean
- Blink duration sd
- Blink quantity
- Blinks total duration
- Pupil diameter mean
- Pupil diameter sd
- Pupil diameter median
- Pupil diameter min
- Pupil diameter max
- Pupil diameter range
- Pupil diameter kurtosis
- Pupil diameter skewness

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fpsyg.2021.653021

Musical and Bodily Predictors of Mental Effort in String Quartet Music: An Ecological Pupillometry Study of Performers and Listeners

Laura Bishop[1,2]*, *Alexander Refsum Jensenius*[1,2] *and Bruno Laeng*[1,3]

[1] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Oslo, Norway, [2] Department of Musicology, University of Oslo, Oslo, Norway, [3] Department of Psychology, University of Oslo, Oslo, Norway

Music performance can be cognitively and physically demanding. These demands vary across the course of a performance as the content of the music changes. More demanding passages require performers to focus their attention more intensity, or expend greater "mental effort." To date, it remains unclear what effect different cognitive-motor demands have on performers' mental effort. It is likewise unclear how fluctuations in mental effort compare between performers and perceivers of the same music. We used pupillometry to examine the effects of different cognitive-motor demands on the mental effort used by performers and perceivers of classical string quartet music. We collected pupillometry, motion capture, and audio-video recordings of a string quartet as they performed a rehearsal and concert (for live audience) in our lab. We then collected pupillometry data from a remote sample of musically-trained listeners, who heard the audio recordings (without video) that we captured during the concert. We used a modelling approach to assess the effects of performers' bodily effort (head and arm motion; sound level; performers' ratings of technical difficulty), musical complexity (performers' ratings of harmonic complexity; a score-based measure of harmonic tension), and expressive difficulty (performers' ratings of expressive difficulty) on performers' and listeners' pupil diameters. Our results show stimulating effects of bodily effort and expressive difficulty on performers' pupil diameters, and stimulating effects of expressive difficulty on listeners' pupil diameters. We also observed negative effects of musical complexity on both performers and listeners, and negative effects of performers' bodily effort on listeners, which we suggest may reflect the complex relationships that these features share with other aspects of musical structure. Looking across the concert, we found that both of the quartet violinists (who exchanged places halfway through the concert) showed more dilated pupils during their turns as 1st violinist than when playing as 2nd violinist, suggesting that they experienced greater arousal when "leading" the quartet in the 1st violin role. This study shows how eye tracking and motion capture technologies can be used in combination in an ecological setting to investigate cognitive processing in music performance.

Keywords: **pupillometry, mental effort, music performance, music listening, musical expression, arousal**

OPEN ACCESS

Edited by:
Steven Matthew Thurman, US Army Research Laboratory, United States

Reviewed by:
Andrea Orlandi, Sapienza University of Rome, Italy
John Lindstedt, SUNY Oswego, United States

***Correspondence:**
Laura Bishop
laura.bishop@imv.uio.no

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *13 January 2021*
Accepted: *17 May 2021*
Published: *28 June 2021*

Citation:
Bishop L, Jensenius AR and Laeng B (2021) Musical and Bodily Predictors of Mental Effort in String Quartet Music: An Ecological Pupillometry Study of Performers and Listeners. Front. Psychol. 12:653021. doi: 10.3389/fpsyg.2021.653021

1. INTRODUCTION

Music performance is a cognitively demanding activity that requires many processes to be carried out in parallel, including overt motor production, covert processing of musical information, monitoring of musical output, and monitoring of audience responses (Bishop and Keller, 2021)[1]. There are additional demands during ensemble performance; for example, performers must divide attention between their own playing and their co-performers' playing (Keller, 2001).

A hallmark of skilled performance is the ability to manage cognitive resources effectively so that accuracy, expressivity, and (in ensemble settings) coordination are maintained. Performances by skilled musicians may seem rather effortless to audience members, but they actually draw on a combination of effortful and automatic processes. These processes involve performers' anticipation of each other's playing (including more effortful imagery and simulation and more automatic melodic expectancies), adaptation to each other's playing (including more effortful period correction and more automatic phase correction), and control of attention (including more effortful directed listening and more automatic passive monitoring).

Skilled performers are able to prioritize one process over another, focusing attention on the prioritized process while non-prioritized processes run automatically. To facilitate attention regulation, musicians may identify landmarks in the music that can serve as cues to attention, drawing their focus to specific technical or expressive processes (Chaffin and Logan, 2006; Chaffin et al., 2010). Less skilled performers may lack attention regulation abilities and, as a result, distribute attention non-optimally—for instance, by focusing on their own playing when they should be listening to their co-performers, or sacrificing expressivity to focus on note accuracy.

In this study, we examine the relationships between the cognitive-motor demands of string quartet performance and attention fluctuations in both performers (Experiment 1) and listeners (Experiment 2). We used the psychophysiological method of pupillometry to gauge changes in the intensity of attention (Kahneman, 1973; Laeng and Alnæs, 2019). In Experiment 1, we collected pupillometry, motion capture, and audio-video data from a string quartet as they performed selections of classical repertoire in our lab in rehearsal and concert/exam conditions. In Experiment 2, we collected pupil data from musically-trained listeners as they heard recordings of the quartet's concert performance. We then analysed how changes in performers' and listeners' pupil diameters related to features of the performers' physical performance (e.g., quantity of motion) and features of the music (e.g., tonal tension and expressivity). In the sections below, we develop some predictions for how these features draw on attention.

1.1. Pupil Size as an Index of Mental Effort

Pupil size is commonly used as an index of attention and mental effort in cognitive tasks (van der Wel and van Steenbergen, 2018; Laeng and Alnæs, 2019). Pupil size is tightly coupled to the release of norepinephrine by the locus coeruleus, which modulates attention and cognitive arousal (e.g., Sara, 2009; Alnæs et al., 2014; Joshi et al., 2016). Pupil dilations occur reliably as part of an orienting response to salient, attention-grabbing stimuli across modalities (Murphy et al., 2016; Marois et al., 2018) or whenever an individual is focused on a challenging task (Laeng et al., 2011). These dilations, described as psychosensory pupil responses (Mathôt, 2018), can be sampled at a fine resolution with modern eye-trackers and act as a gauge of moment-to-moment attention fluctuations.

A number of factors may contribute to how intensely attention is focused at any given moment during a music performance, including the complexity of the music and how technically difficult it is to play. This intensity of cognitive processing is referred to as "mental effort" regardless of the type of task being performed (Kahneman, 1973). Studies of mental effort have shown that pupil dilations occur during complex tasks such as comprehending sentences with higher linguistic complexity (Just et al., 2003), during tasks carried out under interference (O'Shea and Moran, 2019), and when working memory load is high (Kahneman and Beatty, 1966; Klingner et al., 2011; see also Zekveld et al., 2018). Conversely, pupil constriction occurs during periods of distraction and mind wandering (Konishi et al., 2017). Individual differences in cognitive abilities such as working memory capacity also contribute to attention control (Unsworth and Robison, 2017; Endestad et al., 2020).

In a musical context, patterns of pupil dilations reflect listeners' entrainment with musical rhythms (Fink et al., 2018) and listeners' attention to deviations from strict rhythmic regularity. These deviations are referred to as "microtiming" in the context of groove-based jazz music (Skaansar et al., 2019), but are also a common feature of expressively-performed music in many traditions. A pupil response is also observed when listeners hear pitches that deviate from an established tonal context, or in general, when they are surprising (Liao et al., 2016; Bianco et al., 2020). Musical features that capture attention tend to do so reliably across listeners who are familiar with the musical tradition, and as a result, similar patterns of pupil dilation occur among listeners who hear the same musical material (Kang and Wheatley, 2015; Kang and Banaji, 2020).

1.2. Mental Effort and Musical Complexity

Music performance and listening involves continual processing of tonal and timing information (Huron, 2006). Music that violates listeners' expectations for tonality or timing can trigger an increase in mental effort. Fluctuations in mental effort might also occur in response to the complexity of the music that is performed. Musical complexity can be described as a property of a musical stimulus that increases as the degree of uncertainty or unpredictability of pitch, timing, and other features increases. For example, a piece in which many pitch classes have an equal probability of occurring could be deemed more complex than a piece in which few pitch classes are more probable. The complexity of a musical stimulus can also be said to relate to the amount of change that occurs over time (Mauch and Levy, 2011) or the number of events per part or layer.

[1]Bishop, L., and Keller, P. (forthcoming). "Instrumental ensembles," in *The Oxford Handbook of Music Performance*, ed G. McPherson.

From a psychological perspective, it is important to account for how listeners perceive music when assessing its complexity. Listeners' perceptions of complexity are influenced by their long-term musical knowledge, which develops through their exposure to different kinds of music. Studies have shown an inverted U-shaped relationship between complexity and preference (Burke and Gridley, 1990; Gordon and Gridley, 2013), and a relationship between working memory capacity and preference for complexity, mediated by musical training (Vuvan et al., 2020). Marin and Leder (2013) found that arousal mediated a relationship between musical complexity and listeners' ratings of pleasantness.

For the present study, we estimated harmonic complexity, a subcomponent of musical complexity, using a measure of harmonic tension. Harmonic tension and complexity are overlapping phenomena. Both are aggregate constructs that draw on a combination of psychoacoustic features including tonal, temporal, and timbral information. We chose to focus on the tonal component because tonality is particularly relevant in the repertoire that was performed by our quartet. In Western tonal music, moment-to-moment changes in harmonic complexity contribute to listeners' perceptions of harmonic tension, which is usually defined in qualitative terms, with increasing tension described as a feeling of rising intensity and decreasing tension described as a feeling of resolution.

One of our measures of harmonic complexity, "Cloud diameter," derives from the spiral array model of tonality that was proposed for tonal music by Chew (2000). This model is a 3D extension of the circle of fifths, in which pitch classes that are tonally close (e.g., a perfect fifth) are in close spatial proximity to each other. Cloud diameter is computed from musical scores. It represents the tonal distance within a cluster of notes and is given in terms of the spatial distance between their pitch classes in the spiral array (Herremans and Chew, 2016). We predicted that increased Cloud diameter (i.e., increased dissonance) would require increased mental effort to process, resulting in greater pupil size among performers and listeners.

We also obtained ratings of changes in harmonic complexity throughout the pieces from the quartet, which they gave individually per bar for their own parts. These ratings were expected to correlate moderately with Cloud diameter, and were predicted to relate to increased pupil diameter in both performers and listeners.

1.3. Bodily Effort and Mental Effort

Playing music is physically effortful. The processes of carrying out and controlling body motion involves some mental effort. It is therefore important to consider bodily effort when assessing mental effort in performers. In sports, pupil dilations have been shown to occur at the onset of "quiet eye"—the prolonged fixation that expert athletes make on a target immediately prior to initiating a goal-directed action—suggesting heightened mental effort is involved in action preparation (Vickers, 2009; Campbell et al., 2019; Piras et al., 2020). However, the relationship between sustained bodily effort and mental effort generally remains unclear.

For musicians, the physical demands of performance can be described in terms of two components: physical force (or exertion) and control. Force is an important dimension of bodily effort for acoustic instrument performers, as it is the primary means of controlling sound intensity (Olsen and Dean, 2016). Control is central to effective playing technique and relates to how precisely a performer can achieve their intended timing, intonation, timbre, and dynamic level (Palmer, 1997; Bishop and Goebl, 2017). These components of bodily effort can vary independently; for example, performing rapid notes at a low dynamic level requires little force but high control, and can be technically demanding.

In a study by Zénon et al. (2014), pupil size related to the intensity of bodily effort (the amount of force exerted in a power grip task) as well as to participants' perceptions of effort. van der Wel and van Steenbergen (2018) additionally argue that task demands and the amount of effort that participants actually invest in a task can diverge. A recent study of mental effort in imagined and overt piano playing showed a divergence between task demands and measures of mental effort when the difficulty of the task exceeded the capacities of the participants (e.g., when complex movements had to be imagined at a fast tempo; or when imagery had to be carried out with interference; O'Shea and Moran, 2019). Interestingly, even though pupils constricted during the most difficult conditions in this study, participants reported increased mental effort, indicating a dissociation between pupil size and perceptions of effort.

In contrast, a recent study by Endestad et al. (2020), also using pupillometry, showed a clear relationship between pupil size and degrees of mental effort during overt and imagined piano playing for a professional pianist as well as for listeners. This study also showed, using fMRI, differences in locus ceruleous activity for the same pianist as she played (in the scanner) two pieces of different difficulty.

The bodily effort that is involved in playing music affects listeners' experiences of the music as well as performers'. The embodied music cognition framework posits that music perception is a body-based cognitive process that draws on listeners' motor systems in various ways (Maes et al., 2014). Some supporting evidence comes from neuroimaging studies, which have shown that listening to music activates motor circuits, even when the listener makes no overt motion (e.g., Abrams et al., 2013; Gordon et al., 2018). The patterns of activity across motor regions may be especially similar between performers and listeners when they share instrument-specific expertise (Haueisen and Knösche, 2001; Taylor and Witt, 2014).

The activation of motor circuits during music listening may allow listeners to covertly simulate features of the actions that were involved in playing the music (Wilson and Knoblich, 2005; Repp and Knoblich, 2007). This real-time simulation of music performance actions may help listeners to generate predictions that shape their perception of the music. Motor activation may also occur at a more general level, allowing for simulation of features of actions that listeners have no experience in performing, or remapping of action features to familiar action sequences (e.g., allowing a listener to covertly sing along with a melody played by a violin, despite having no violin-playing

experience; Godøy et al., 2005; Eitan and Timmers, 2010; Maes et al., 2014; Kelkar and Jensenius, 2018). Motor activation during music listening may furthermore help listeners to construct expressive interpretations of the music they hear and relate to the performer(s) on an emotional level (Molnar-Szakacs et al., 2011; Olsen and Dean, 2016).

In the current study, we assessed the effect of bodily effort on pupil changes during string quartet performance and listening. We predicted that the bodily effort that performers invested in their playing would engage increased mental effort. We also predicted that the bodily effort that listeners heard in recordings of the quartet's performances would engage increased mental effort, especially in the case of string musicians, who would be most familiar with the sound-producing actions involved in quartet playing. Bodily effort was operationalized in terms of several different measures, which we selected to capture different aspects of the physical demands of playing a stringed instrument. These included measures of overt head and arm motion, acoustic intensity (taken as a correlate of physical force), and the performers' subjective, per-bar ratings of technical difficulty.

1.4. Expressivity, Arousal, and Mental Effort

Attention-related modulations of pupil size also reflect changes in arousal. For both performing musicians and listeners, local fluctuations in arousal can occur in relation to expressive changes in the music (Schubert, 2004; Lundqvist et al., 2009; Egermann et al., 2015). Musical expression is a construct that arises from interactions between different musical parameters, including pitch, timing, dynamics, timbre, and various body features, among others (Juslin, 2003; Jensenius et al., 2010; Cancino-Chacón et al., 2017). For music in the Western classical tradition, expressivity is to a large extent tied to certain key structural features that are given in a score (Palmer, 1997). Performers may interpret these features in different ways, thus producing performances that are expressively distinct.

One component of musical expression is emotional expression (Juslin, 2003). Music can be emotionally expressive in different ways, including through extramusical associations and through perceptual expectations that arise from familiar harmonic relationships (Egermann et al., 2013; Pearce, 2018). The emotional qualities of music are commonly described in terms of arousal and valence (e.g., Schubert, 2004). Many of the basic emotions that people report associating with music can be readily placed in a two-dimensional space that crosses arousal with valence (e.g., happiness, peacefulness, fear, etc.), though this is not the case for some more complex emotions, such as being moved (*kama muta*) or awe, which contain elements of seemingly contradictory emotional states (e.g., awe is described as including aspects of both sadness and joy; Konecni, 2005; Menninghaus et al., 2015; Zickfeld et al., 2019).

A few studies have investigated the relationship between pupil size and emotional arousal in either music performers or listeners. Gingras et al. (2015) showed a positive relationship between pupil dilation and ratings of emotional arousal and tension in listeners who heard brief (6-s) excerpts of Romantic-style piano trios. In another study, pupil dilations were shown to occur in close temporal proximity to listeners' reported chills (i.e., peak emotional experiences; Laeng et al., 2016). In a study investigating the arousal elicited by vocal vs. instrumental melodies, listeners displayed a more dilated pupil when hearing vocal melodies than when hearing the same melodies played on a piano, suggesting that the human voice is treated as a "privileged signal" (Weiss et al., 2016). In the same study, listeners also showed a more dilated pupil when hearing familiar melodies than when hearing novel melodies.

The current study contributes to this literature with an investigation of how local fluctuations in expressivity relate to pupil size. Expressivity was quantified through performers' per-bar ratings of expressive/interpretive difficulty. It should be noted that we did not ask for performers' ratings of music-related arousal or expressive intensity, although we expect that these factored into the ratings that they gave (see Methods for our exact wording). Their ratings may also reflect their judgements of musical complexity and technical difficulty, which also contribute to how readily performers realize their expressive goals. In short, our measure of expressive difficulty probably constitutes a higher-order indication of the performers' relationships with the music. We predicted that higher ratings of expressive difficulty would correspond to higher emotional arousal and, correspondingly, larger pupil size for both performers and listeners.

1.5. Arousal and Attention Regulation During Music Performance

Changes in performers' arousal can occur across relatively long timeframes (e.g., across the course of a concert). These changes occur in addition to the local fluctuations that relate to musical expression, and can be assessed with pre-trial "baseline" pupil measurements. For performers, baseline levels of arousal are likely to depend on the conditions surrounding their performance (e.g., who is in the audience, how well-prepared the performers are, etc.) and their individual response to those conditions. Elevated arousal prior to public performance is common, and often associated with performance anxiety (Kenny, 2011). Performances are given optimally under moderate levels of arousal (Papageorgi et al., 2007). Physiological correlates of autonomic arousal, including increased heart rate, increased motor excitability, and sweating, can themselves be detrimental to performance, impairing fine motor control and increasing the bodily effort that is required to maintain technical accuracy. Musicians may have to deviate from their practiced playing technique in order to compensate and maintain control of their movements (Yoshie et al., 2009). This makes performance more difficult and adds to musicians' mental workload. Absorption (sometimes described in terms of flow) is noted to emerge predominately under moderate levels of arousal (Peifer et al., 2014; Vroegh, 2019).

Effective attention regulation is also thought to require a moderate level of arousal (Unsworth and Robison, 2017). Lenartowicz et al. (2013) suggest that high and low levels of arousal pave the way for different types of distractibility. They posit a "landscape" of attention control states based on crossing high and low arousal levels with internal and external

attention orientations. If arousal is low, and internal focus can lead to mind wandering and zoning out, while an external focus leads to behaviour comprising predominately automatic responses to salient stimuli. If arousal is high, an inwards focus can result in mind-racing, while an external focus leads to excessive and nondiscriminating responses to both relevant and irrelevant stimuli.

For ensembles like a classical string quartet, baseline arousal levels might be expected to differ between performers according to their individual roles in the ensemble. Traditionally, the 1st violinist is the leader of the group. This is particularly the case for repertoire from the classical period (e.g., including works by Haydn and Mozart), where the 2nd violinist, violist, and cellist typically have more supporting roles. The 1st violinist is also often responsible for giving cues to the other musicians to help keep the group together. While string quartets may operate more democratically in other ways (e.g., jointly making interpretative decision), anecdotal evidence suggests that some 1st violinists feel heightened stress in performance due to their leadership role (Davidson and Good, 2002). If this is heightened stress occurs, it is likely reflected in the 1st violinist's pupil dilations. We predicted that the 1st violinist would show more dilated pupils than the other musicians. We also predicted that the musicians would show more dilated pupils in the baseline measurement taken just before the start of the concert performance than in the baseline measurements taken before the rehearsal performances, earlier in the same recording session.

1.6. Current Study

This study made a novel assessment of how musical complexity, bodily effort, expressive difficulty, and situational factors (including rehearsal vs. concert setting, piece order, and musical role) contribute to mental effort in performers and listeners of string quartet music. In Experiment 1, we invited a student string quartet from a local music academy to record some performances of their current repertoire at our lab. The quartet gave five performances of an excerpt from one of their pieces in rehearsal conditions (i.e., with no audience). We manipulated the configuration of the quartet across rehearsal performances in order to partially or completely disrupt visual communication between musicians. Bishop et al. (2021)[2] reports on how these manipulations affected interperformer communication during the rehearsal performances. The current paper does not consider these manipulations further.

Following the rehearsal performances, the quartet played the full set of pieces for a live audience (which included an examiner) in a concert/exam condition. We collected pupillometry, gaze, motion capture, and audio/video data from the musicians. The performers later individually provided per-bar ratings of Harmonic complexity, Technical difficulty, and Expressive difficulty. In Experiment 2, we collected pupillometry data from 16 trained musicians as they listened to recordings of the quartet's concert performance.

[2]Bishop, L., Gonzalez Sanchez, V., Laeng, B., Jensenius, A. R., and Høffding, S. (submitted). Variability of head motion and gaze across social contexts in string quartet performance.

Musical complexity, bodily effort, and expressive difficulty were subdivided into a combination of predictors that included ratings provided by the performers and measurements of score information and performance data (**Figure 1**). With this combination of predictors, we aimed to capture the effects of perceived effort, allocated effort, and task difficulty. The overarching prediction was that subjective and objective measures of both musical complexity and bodily effort would relate to increased pupil dilations. We predicted a similar pattern of results for performers and listeners, although we expected a larger effect of predictors relating to performers' bodily effort on performers than on listeners.

As predictors relating to musical complexity, we included Cloud diameter as a measure of harmonic tension, and performers' ratings of Harmonic complexity. As predictors relating to bodily effort, we included quantity of head and arm motion, energy (acoustic intensity) of the musical sound signal, and performers' ratings of technical difficulty. For string musicians, head motion is not directly involved in sound production, but may be representative of musicians' expressive engagement with the music (see Glowinski et al., 2013a,b) and communication with co-performers (Bishop and Goebl, 2018)[2]. Sound intensity can be considered a correlate of physical force, as more forceful movement is required to produce higher-intensity audio signals on string instruments. Our measure of expressive difficulty, which we posited incorporated aspects of arousal, musical complexity, and technical difficulty, constituted ratings provided by the performers.

We additionally made several predictions relating to situational factors, which were tested in Experiment 1. First, we predicted that arousal would be greater at the start of the concert than at the start of the rehearsal period. We also predicted that during the course of the concert, some global changes in performers' tonic arousal would occur as their initial anxiety reduces. We tested for differences in mean pupil size between the four pieces that the quartet played in the concert, expecting that a gradual decline in arousal would occur. We also predicted that levels of arousal would be tied to the different musical roles of the quartet members. The violinists switched roles halfway through the concert, so that each played the 1st violin part for two of the four pieces. This gave us an opportunity to test the prediction that the 1st violinist would have heightened arousal due to their leadership role.

2. EXPERIMENT 1: MENTAL EFFORT IN STRING QUARTET REHEARSAL AND CONCERT PERFORMANCE

2.1. Participants

A student string quartet from a local music academy took part in the experiment (1 female, 3 males; ages 19–20; 13–16 years of music training). They had established themselves as a group 6 months prior to the experiment, but had occasionally played together in various ensembles before this, having attended the same music school as children. The 2nd

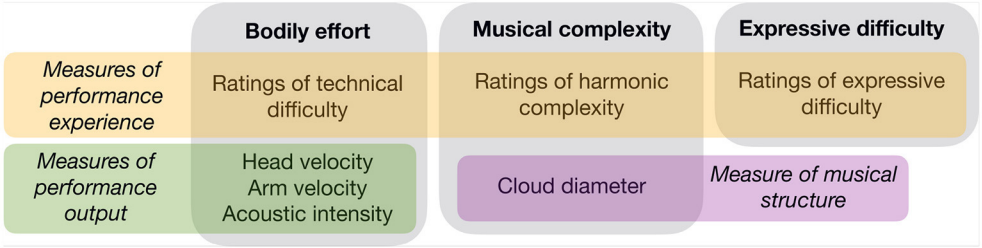


FIGURE 1 | List of predictors relating to bodily effort, musical complexity, and expressive difficulty.

violinist and cellist had played together in a quartet for 11 years. At the time of the experiment, the quartet had been rehearsing the Haydn piece for 3 months and the Debussy piece for 2 months. The musicians all provided written informed consent.

2.2. Materials and Equipment

Motion data were recorded using a Qualisys system with 12 Oqus 300 cameras (performances 1–6) and an OptiTrack system with 8 Flex 13 cameras (performance 7); see **Figure 2**[3]. The musicians wore jackets and caps with six reflective markers attached (1 on the head, 1 on the upper back, and 2 on each arm above and below the elbow). Marker positions were sampled at 120 Hz. Pupil and gaze data were collected using SMI Wireless Eye Tracking Glasses, which recorded at 60 Hz. To synchronize motion and eye data, we recorded an audiovisual signal using a clapperboard with 2 reflective markers affixed at the start of each performance. Recordings were aligned retrospectively from this point. Additional details on our equipment set-up relating to measures that are not reported here are given by[2].

The experiment was carried out in our lab, which has moderately bright lighting and black curtains covering the walls and windows. To avoid adding noise to the pupillometry data, we did not use any spotlights or stage lighting.

During the rehearsal performances, the musicians played the first 68 bars of the first movement of the String Quartet in B-flat major, Op. 76, No. 4, by Haydn. For the concert, they played the full first and second movements of this work as well as the full first and second movements of the String Quartet in G minor, Op. 10, by Debussy. Hereafter, we will refer to these pieces as Haydn I and II and Debussy I and II.

2.3. Procedure

The musicians warmed up briefly and were then positioned for the first performance. They completed a 1-min baseline pupil recording before playing, for which they were instructed to sit still and focus their gaze on a single score note. The performance was then recorded, and the musicians were repositioned for the next condition. Baseline recordings were made every time

FIGURE 2 | Photos showing the eye-tracking glasses and locations of body markers on the musicians. Photo credit: Annica Thomsson.

the musicians were repositioned. Following the Replication-rehearsal, we paused so that the musicians could take a break and the lab could be set up with audience seating for the concert. A final baseline and then the concert were recorded. In total, including setup and breaks, the experiment took around 4 h.

In the weeks following the recording session, the musicians made per-bar ratings of perceived difficulty for the music that they played during the concert. They rated the music on three measures, using a scale of 1—7: technical difficulty (how technically challenging was the bar to play?); expressive/interpretative difficulty (how difficult was it to express the intended meaning, idea, or emotion?); and harmonic tension/complexity (how harmonically complex or tense was the material in the bar?). The rating task was done separately and individually and the musicians submitted scanned and rated copies of their scores when they had finished.

2.4. Analysis

2.4.1. Preprocessing of Pupil, Motion, and Musical Data

2.4.1.1. Pupil Data

We used binocular pupil diameters (in mm) for our analysis, which we obtained by averaging the values that were recorded for left and right eyes. We used binocular rather than monocular values in order to minimize any effects of any outliers that might occur in one eye or the other (especially in moments where participants were looking at more extreme angles). A multi-step procedure was developed for cleaning and filtering binocular pupil data. First, to eliminate blinks, we discarded any observations where the recorded diameter was more than 2 standard deviations below the mean diameter for the trial. We

[3]Both systems actually recorded the full experiment, and (Bishop and Jensenius, 2020) shows that they were comparable in recording quality. We used OptiTrack data for performance 7 because the Qualisys recording was started a few seconds late.

found that it was also necessary to discard some non-zero values at the edges of the "gaps" that occurred because of blinks, where the eye was captured partially closed. These values were identified on the basis of velocity, which was calculated as the first derivative (rate of change) of the series of pupil diameters. Observations where the velocity of diameter change was more than 2 standard deviations from the mean velocity of the trial were discarded. A Savitzky-Golay filter was used to smooth the remaining data (order = 3, window = 11), and gaps in the data were filled using a linear interpolation. Finally, blinks were removed from baseline data, and smoothed performance data were calculated as differences from mean baseline diameters.

Since the ratings provided by the performers were given per bar, we downsampled the processed pupil data to obtain an average diameter per bar. To do this, we interpolated a series of bar onset times using the audio recordings. Bars were assumed to be evenly spaced in time. Although we acknowledge that this introduces some imprecision into our alignments, a more precise audio-to-score mapping would have been a substantial task and beyond the scope of this project. Pupil data and bar numbers were then aligned based on their timestamps, and we calculated an average pupil diameter per bar. These averaged, per-bar diameters were analysed in the linear mixed effects models (see below).

2.4.1.2. Motion Data

Head and arm data were used for the analyses presented here. We chose to focus on *velocity* rather than a higher order kinematic feature (e.g., smoothness) because velocity provides a more direct measure of the mechanical energy that is expended by a performer and is related to momentum. Smoothed velocities were derived using a Savitzky-Golay filter (order = 3, window = 41; "savitzkyGolay" function from the "prospectr" package in R, which optionally outputs smoothed derivatives of the input data). The norm of smoothed 3D velocities was then computed. Using the bar onset times that we describe above (see Pupil data), we aligned the motion data with bar numbers based on their timestamps. "Quantity of motion" (QoM) was then calculated as the sum of velocities per bar of each piece.

2.4.1.3. Audio Data

Root mean square (RMS) values were extracted from audio recordings as a measure of acoustic intensity. This was done in Python using the package Madmom (Böck et al., 2016), with a frame size of 2048 samples and 50% overlap. RMS curves were smoothed using a convolution-based method, with a Hamming window of 50 samples. The resulting RMS values were averaged per bar for the linear mixed effects model analysis. Hereafter we refer to these values as "sound level."

2.4.1.4. Cloud Diameter

Cloud diameters were calculated for the score of each piece in Python, using the package Partitura (Grachten et al., 2019), at increments of 1 bar. The algorithm requires scores in musicXML format. We obtained MIDI files for all of the pieces online[4], hand-corrected them for pitch spelling (with reference to the scores that were used by the quartet), and converted them to musicXML in MuseScore[5]. Output Cloud diameters are given in units of a perfect fifth in Chew (2000)'s spiral array.

2.4.2. Linear Mixed Effects Modelling of Pupil Diameter

Linear mixed effects models (LMMs) were used to test the contribution of predictors relating to musical complexity, bodily effort, and expressive difficulty to baseline-normalized pupil diameter. This was done using the "glmmTRB" package in R.

We tested two models, one which included head motion as an index of bodily effort, and one that included arm motion instead. Model 2 (with arm motion) included fewer data points than Model 1 (with head motion) because some of the arm markers, especially for the 2nd violinist and violist, were not as well tracked as the head markers were. As a result, we could only include arm data for 1–2 pieces for these performers. Nonetheless, it was important to consider arm motion as a measure of bodily effort because it is directly tied to sound production.

- **Model 1** included seven fixed effects: quantity of Head motion, Sound level, Technical difficulty ratings, Cloud diameter, Harmonic complexity ratings, and Expressive difficulty ratings.
- **Model 2** included quantity of Arm motion (*instead of* Head motion), Sound level, Technical difficulty ratings, Cloud diameter, Harmonic complexity ratings, and Expressive difficulty ratings.

For both models, musician and performance were included as crossed random effects. Since our predicted variable constituted time series data, we also specified an autocorrelation structure (order = 1) with time (in bars) as a covariate and the same grouping structure as our random effects.

The formulation of Model 1 was as follows:

$$\textit{Pupil size} \sim \textit{Cloud diameter} + \textit{Harmonic complexity ratings} + \textit{Head motion} + \textit{Sound level} + \textit{Technical difficulty ratings} + \textit{Expressive difficulty ratings} + (1|\text{piece}) + (1|\text{ID}) + \text{ar1}(\text{bars} + 0|\text{piece:ID})$$

The formulation of Model 2 was as follows:

$$\textit{Pupil size} \sim \textit{Cloud diameter} + \textit{Harmonic complexity ratings} + \textit{Arm motion} + \textit{Sound level} + \textit{Technical difficulty ratings} + \textit{Expressive difficulty ratings} + (1|\text{piece}) + (1|\text{ID}) + \text{ar1}(\text{bars} + 0|\text{piece:ID})$$

To estimate effect sizes, we used a hierarchical modelling procedure in which significant predictors from Models 1 and 2 were added incrementally one by one to a null model containing only the intercept term and random effects. Predictors were entered in decreasing order of absolute estimate size (i.e., in the order they are listed in **Table 1**; see Data Sheet 1 in **Supplementary Material**). These hierarchically constructed models were then compared against the null model. We report χ^2 tests and Bayesian Information Criterion (BIC) values

[4]kunstderfuge.com

[5]musescore.com

as indications of effect size. BIC is commonly used as a criterion for model selection. To protect against overfitting, it incorporates a penalty for the number of predictors that are included in a model. Lower BIC indicates better support for a given model.

2.4.3. Effects of Rehearsal vs. Concert Setting, Musical Role, and Piece/Concert Time

To test for differences in levels of baseline arousal between the start of the rehearsal and the start of the concert, we compared the pupil diameters that were recorded in the first rehearsal baseline with the pupil diameters that were recorded in the baseline before the concert performance, using a Wilcoxon Signed Rank test.

We also compared average pupil diameters between the four concert pieces for each performer individually. Only data from the concert were used in this part of the analysis. Series of LMMs were run for each performer individually that included concert piece as a fixed effect. Concert piece was also included as a random effect in each model, and we specified an autocorrelation structure (order = 1) with time in concert piece as a covariate. We ran three models per performer with different concert pieces set as the base level for contrasts, so that we could get the full set of between-piece contrasts (6 total). We tested for significance at $\alpha = .008$, following Bonferroni adjustment. Importantly, the "effect of concert piece" that we tested with these models reflects not only differences in musical material, but also the passing of concert time, and in some cases changes in musical role (the violinists exchanged places after Haydn II).

2.5. Results

2.5.1. Linear Mixed Effects Modelling of Pupil Diameter

The reader is referred to **Figure 1** for a reminder of which predictors we tested. The results of the LMMs are given in **Table 1**. Model 1 showed positive effects of Technical difficulty and Expressive difficulty on pupil size, and negative effects of Cloud diameter and Harmonic complexity. Head motion and Sound level did not yield significant effects. Models containing the four significant predictors were compared against a null model, and all predictors yielded significant χ^2 values. However, only the model containing Technical difficulty and the model containing Technical difficulty and Cloud diameter improved the BIC, suggesting that the effects of Harmonic complexity and Expressive difficulty on pupil diameter were weak.

Model 2 showed positive effects of Arm motion and Technical difficulty on pupil size, and negative effects of Cloud diameter and Harmonic complexity. Sound level and Expressive difficulty did not yield significant effects. When we compared a null model against a hierarchical series of models containing the four significant predictors, all predictors yielded significant χ^2 values and reduced the BIC relative to the null model.

In summary, both models showed stimulating effects of bodily effort (Technical difficulty, Arm motion) and negative effects of musical complexity (Harmonic complexity, Cloud diameter). Only Model 1 showed a stimulating effect of Expressive difficulty. Descriptive plots for significant predictors are given in the **Supplementary Figures 1–5**.

TABLE 1 | Results of linear mixed effects modelling for performers.

Model	Fixed effect	Estimate	SE	*z*-value	χ^2	BIC
Model 1	(null model)	—	—	—	—	–1240.7
	Technical difficulty	0.0316	0.0038	8.21***	102.37***	–1334.8
	Cloud diameter	–0.0192	0.0029	6.56***	51.74***	–1378.4
	Harmonic complexity	–0.0120	0.0041	2.95**	6.88**	–1377.0
	Expressive difficulty	0.0098	0.0039	2.51*	5.42*	–1374.2
	QoM head	0.0003	0.0003	1.24	—	—
	Sound level	1.4e-5	1.4e-5	1.05	—	—
Model 2	(null model)	—	—	—	—	–907.40
	Technical difficulty	0.0216	0.0045	4.80***	33.53***	–933.10
	Harmonic complexity	–0.0188	0.0049	3.81***	21.22***	–946.48
	Cloud diameter	–0.0159	0.0033	4.80***	19.63***	–958.28
	QoM arms	0.0009	0.0001	6.49***	43.68***	–994.12
	Expressive difficulty	0.0057	0.0046	1.25	—	—
	Sound level	2.2e-5	1.59e-5	1.39	—	—

*Predictors are listed in descending order of absolute estimate size. Negative estimates indicate predictors that had a negative effect on pupil diameter. SE indicates standard error. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

TABLE 2 | Correlations between performers' rating measures.

	Harmonic complexity	Technical difficulty
Technical difficulty	0.43*	—
Expressive difficulty	0.41*	0.58*

*$p < 0.001$.

TABLE 3 | Correlations between measures relating to bodily effort.

	Technical difficulty ratings	Sound level
Head motion	–0.001	0.31*
Arm motion	0.02	0.41*
Sound level	0.08*	—

*$p < 0.001$.

We also evaluated the similarity between predictors that could be expected to overlap. **Table 2** lists the correlations between performers' rating measures. **Table 3** lists the correlations between measures relating to bodily effort. The correlation between musical complexity measures (Harmonic complexity ratings and Cloud diameter) was slight, $r = .28, p < .001$.

2.5.2. Effects of Rehearsal vs. Concert Setting, Musical Role, and Piece/Concert Time

Figure 3 shows mean pupil diameters for each performer/piece combination. Our comparison of baseline pupil size captured before the first rehearsal with baseline pupil size captured before the first concert performances showed no significant difference, $M = 3.72$ mm, $SD = 1.05$ mm (rehearsal); $M = 4.08$ mm, $SD = .96$ mm (concert); $W = 4, p = .34$.

Results of the LMMs testing within-performer/between-piece difference in pupil diameter are shown in **Table 4**. For the 1st violinist, pupil dilation was greatest in the first piece (Haydn I). For the 2nd violinist, pupil dilation was greater in the Debussy

TABLE 4 | Results of the LMMs testing within-performer/between-piece differences in pupil diameter.

Performer	Contrast	Estimate	SE	*z*-value
1st Violin	Haydn I vs. Haydn II	–0.3331	0.0566	5.89**
	Haydn I vs. Debussy I	–0.3781	0.0494	7.65**
	Haydn I vs. Debussy II	–0.3327	0.0499	6.67**
	Haydn II vs. Debussy I	–0.0450	0.0564	0.80
	Haydn II vs. Debussy II	0.0004	0.0571	1.00
	Debussy I vs. Debussy II	0.0453	0.0499	0.36
2nd Violin	Haydn I vs. Haydn II	–0.0065	0.0556	0.012
	Haydn I vs. Debussy I	0.2653	0.0469	5.66**
	Haydn I vs. Debussy II	0.3363	0.0475	7.08**
	Haydn II vs. Debussy I	0.2718	0.0554	4.90**
	Haydn II vs. Debussy II	0.3428	0.0560	6.13**
	Debussy I vs. Debussy II	0.0710	0.0474	1.50
Viola	Haydn I vs. Haydn II	0.1322	0.0541	2.44
	Haydn I vs. Debussy I	0.0335	0.0452	.74
	Haydn I vs. Debussy II	0.1547	0.0458	3.38**
	Haydn II vs. Debussy I	–0.0987	0.0537	1.84
	Haydn II vs. Debussy II	0.0225	0.0539	0.42
	Debussy I vs. Debussy II	0.1212	0.0456	2.66*
Cello	Haydn I vs. Haydn II	–0.0460	0.0505	0.91
	Haydn I vs. Debussy I	0.0828	0.0417	1.99
	Haydn I vs. Debussy II	0.1110	0.0423	2.63
	Haydn II vs. Debussy I	0.1288	0.0505	2.55
	Haydn II vs. Debussy II	0.1569	0.0509	3.08*
	Debussy I vs. Debussy II	0.0281	0.0421	0.67

**$p < 0.001$, *$p < 0.008$.

pieces than in the Haydn pieces. This is notable because the 2nd violinist played as 1st violinist for the Debussy pieces. The violist showed greater pupil dilation in Debussy II than in Haydn I or Debussy I. The cellist showed greater pupil dilation in Debussy II than in Haydn II.

During the quartet's concert performance of Haydn I, an unexpected event occurred: the 1st violinist mishandled a page turn and, as a result, had to play the last page from memory. The sudden uptake in arousal and prolonged increase in mental effort are clear in the timecourse of his pupil diameter curve (**Figure 4**). We would note that this incident does not entirely account for the 1st Violinist's high average pupil diameter during that performance. As can be seen from the plot, the 1st Violinist's pupil was dilated (relative to the other musicians) from the start. We would also note that the 1st violinist's response to the page turn incident had no noticeable effect on the results that are presented in section 2.5.1. Models 1 and 2 were rerun on a data subset that excluded the 1st violinist after the moment of the error and the pattern of significant and nonsignificant effects remained the same.

2.6. Discussion

This experiment evaluated the contributions of musical complexity, bodily effort, and expressive difficulty to pupil size in performing musicians. In this section, we will focus on the effects of bodily effort, musical complexity, and expressive difficulty, which informed our decision to carry out Experiment 2. We will also discuss the effects of situational factors, including within-performer/between-piece differences in pupil size.

2.6.1. Effects of Bodily Effort, Musical Complexity, and Expressive Difficulty on Pupil Size

Both Models 1 and 2 showed negative effects of Harmonic complexity and Cloud diameter on pupil size. We included Cloud diameter in our analysis as a more systematic measure of harmonic complexity to complement performers' subjective ratings. As we explained in the Introduction, Cloud diameter provides an indication of the degree of dissonance in each chord. It is notable that Cloud diameter and ratings of Harmonic complexity were only slightly correlated. The performers may have accounted for other aspects of harmonic complexity in their ratings (e.g., number of distinct tones, or amount of chord-to-chord change). Performers might also have weighted the relative complexity of chords within each bar less systematically than we achieved by calculating per-bar Cloud diameters. Despite the limited overlap between Cloud diameter and Harmonic complexity ratings, both measures yielded similar, unexpectedly negative effects on pupil size.

We had predicted that increased musical complexity would demand more effortful music processing and prompt increased pupil dilation, so this result was unexpected. Harmonic complexity is one component of the broader construct of musical complexity, and may share a complex relationship with other components, such as metric or rhythmic complexity, which also place (potentially competing) demands on attention. Thus, a potentially stimulating effect of Harmonic complexity might have been masked by other musical factors.

Harmonic complexity might also be less relevant to the experience of mental effort in string quartet performance than we originally predicted. Most quartets have spent a lot of time rehearsing by the time they perform in concert, and therefore have a close familiarity with the music. This familiarity might change the way they process the harmonic information that is contained in the music, perhaps reducing the mental effort that processing requires. In order to determine whether harmonic complexity affects performers and listeners differently, we designed Experiment 2 especially with the aim of examining the relationship between harmonic complexity and mental effort in listeners.

Performers' ratings of Technical difficulty had a positive effect on pupil size for both models. Indeed, for both models, Technical difficulty yielded a larger absolute estimate size than the other significant effects. These results are in line with our prediction that increased technical difficulty would engage increased mental effort. The Haydn and Debussy String Quartets are stylistically different and present performers with a variety of challenges. The performers used a large range of technical difficulty ratings for all pieces (1–6 for Haydn I, 1–5 for Haydn II, 1–7 for Debussy I, 1–6 for Debussy II), suggesting that they considered these variety of challenges in their evaluations. It is interesting that perceived technical difficulty seems to have a continued effect on mental effort when the musicians are playing well-practiced music and likely have many aspects of their bodily performance automatized.

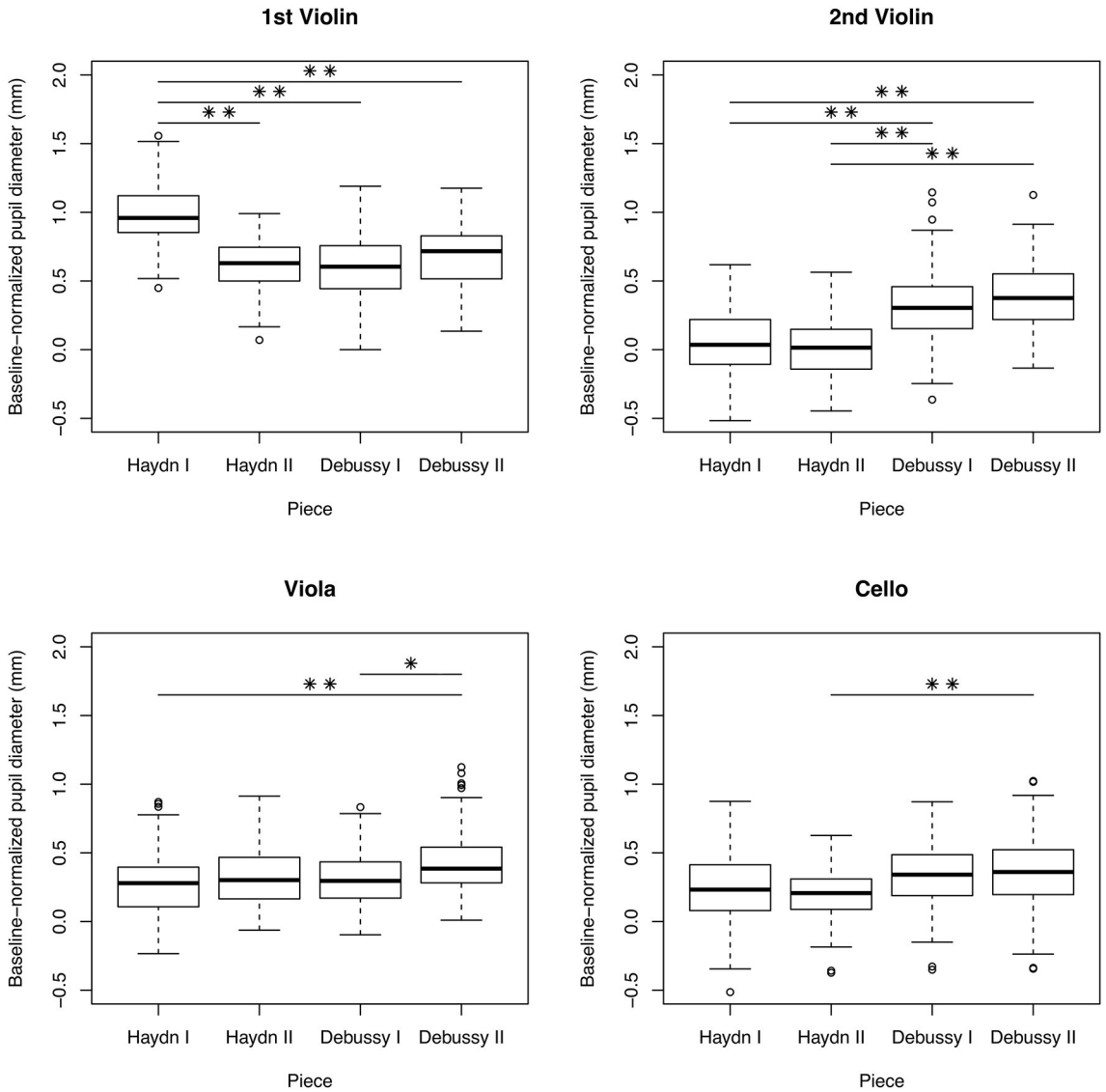


FIGURE 3 | Boxplots showing pupil diameters across concert performances for each musician. Note that the violinists switched places for the Debussy pieces, so the 1st Violin played the 2nd Violin part and vice versa. $^{**}p < 0.001$, $^{*}p < 0.008$.

Model 1 showed no effect of Head motion, while Model 2 showed a positive effect of Arm motion. For string musicians, head motion is primarily expressive. This is in contrast to arm motion, which is more directly involved in sound production. Arm motion happens at a much faster pace and must be controlled at a much finer level than is the case for head motion. In particular, control is needed for carrying out correct fingering and positioning the left hand so as to maintain intonation, as well as for carrying out appropriate bowing technique (Dalmazzo and Ramírez, 2019; D'Amato et al., 2020) and achieving coordination between the two hands. As a result, arm motion likely requires more substantial bodily effort than does head motion, and may have a more arousing effect on the body because it requires more physical exertion.

Model 1 showed a positive effect of Expressive difficulty on pupil size, in line with our prediction, though the effect was weak, and was not significant in Model 2. We presume that expressive difficulty incorporates a combination of demands relating to technical difficulty, complexity, emotional engagement, and perhaps coordination difficulty (which we did not evaluate here; see General Discussion). The positive effect that we observed in Model 1 suggests that this higher-order measure explains some variance in pupil diameter, above and beyond that explained by the other predictors relating to bodily effort and complexity.

In summary, both models showed positive effects of technical difficulty ratings and negative effects of Harmonic complexity ratings and Cloud diameter. The models differed in their results for body motion (quantity of arm motion had a positive effect

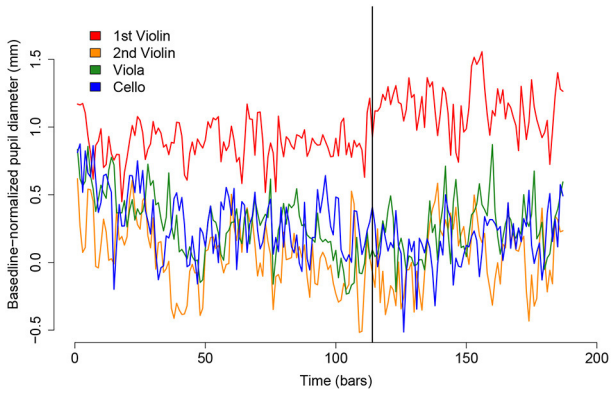


FIGURE 4 | Timecourse of pupil diameters for the quartet across the course of the 1st movement of the Haydn String Quartet. The black vertical line indicates the location of the 1st violinist's failed page turn.

on pupil size; quantity of head motion was nonsignificant) and expressive difficulty. As we have discussed, the difference in effects of head and arm motion likely has to do with the different musical functions that these parts of the body have, as well as the degree of exertion and control that they require. The effect of expressive difficulty in Model 1 was significant but weak. The lack of effect of expressive difficulty in Model 2 might be due to inclusion of arm motion instead of head motion. Perhaps arm motion overlaps with expressive difficulty to a greater extent than does head motion. Model 2 also used a reduced dataset due to the poor tracking of some arm markers, which may have rendered the already-weak effect of expressive difficulty less clear.

2.6.2. Effects of Situational Factors on Pupil Size

While performers demonstrated a slightly larger pupil diameter at the start of the concert than at the start of the rehearsal, this difference was not significant. This lack of effect is in contrast to our prediction that arousal would be higher before the concert. The quartet may have been anxious at the start of the rehearsal, since they had not performed in our lab before, and had to get used to the unusual setting, the motion capture and eye tracking equipment, and the non-optimal acoustics. Therefore, they may have experienced a high baseline arousal at the very beginning of the session, which reduced as they acclimatized to the lab environment.

The lack of effect here reminds us that performance always occurs in the context of some social and material environment (van der Schyff et al., 2018), which unavoidably has some effect on performers' arousal. We should also be wary of blindly categorizing concert and rehearsal performances as high and low arousal. Performers may feel more pressure to perform well under some rehearsal conditions (e.g., when playing in an unfamiliar place, when rehearsing for the last time before an important concert) than in some concert conditions (especially if the concert is relatively low stakes). In future research, studies of arousal during public performance in ecological settings should take into account the performance environment and performers' goals and mindset, to show more clearly how changes in arousal relate to the performers' placement in a specific concert situation. This could be done with a mixed-methods approach that includes physiological/behavioural measures and interviews/questionnaires, similar to the paradigm proposed by Bojner Horwitz et al. (2020).

Our within-subject/between-piece analysis yielded some notable findings. The first violinist exhibited a very dilated pupil during the first piece (Haydn I). Based on our debriefing discussions with the musicians, we understand that he was feeling anxious at the start of the concert. This anxiety was exacerbated by a failed page turn partway through Haydn I, which necessitated him to play the last pages of the piece by memory. Although his pupils remained dilated through the rest of the concert relative to the other quartet members, we did see a significant reduction in his pupil size between Haydn I and Haydn II. The second violinist showed smaller pupil sizes during the Haydn pieces, when he was playing as second violin, than during the Debussy pieces, when he was playing as first violin. Thus, both violinists showed greater arousal when playing as first violin than when playing as second violin. This is in line with the prediction that the first violin leadership role comes with additional demands (Davidson and Good, 2002; Timmers et al., 2014; Glowinski et al., 2015). As we have previously reported, during the rehearsal performances, the first violinist was distinct in his visual attention, and almost never looked at any of the other musicians[2]. In contrast, the other musicians looked at him more than they looked at any other quartet member. Combined with the current results, it seems that the whole quartet recognized the first violinist as the leader, and that this had substantial effects on how everyone interacted with each other.

3. EXPERIMENT 2: MENTAL EFFORT IN MUSIC LISTENING

Experiment 2 was designed to follow up on some of the findings from Experiment 1. Given the small sample size in Experiment 1 (n = 4), we wanted to test whether the effects that we observed there would reemerge in a larger sample. A follow-up experiment with listeners would also allow us to shed some light on how mental effort and arousal compare across performance and listening tasks. An especially interesting question is how much performers' bodily effort contributes to the experiences of musically-trained listeners. Does technical difficulty also demand increased mental effort among listeners? Another interesting question is how strongly performers' subjective ratings of difficulty and complexity contribute to listeners' experiences. While we would expect some widespread agreement on what is complex or difficult, performers differ in their skills and anatomical/physiological constraints (e.g., hand size, strength, ability to move rapidly, etc.), so they necessarily show some variability when rating these factors. If the subjective ratings given by a small sample of four performers contribute significantly to listeners' mental effort, then this will indicate some generalizability to those ratings.

3.1. Participants

Sixteen trained musicians (8 female/8 male) completed the listening task. Five of the musicians were violinists (3), violists (1), or cellists (1), and rest of the musicians played a variety of other instruments (guitar–3, piano–3, flute–1, percussion–1, sitar–1, trombone–1, voice–1). Separate analyses were run for string and non-string musicians, but revealed no between-group differences, so we merged all participants together into a single group. The musicians were on average 26.4 years old ($SD = 5.7$) and had on average 13.3 years of musical training ($SD = 6.7$).

We asked the musicians to rank their familiarity with each piece on a scale of 1-4 (1 = "Never heard it before"; 2 = "I think I've heard it before"; 3 = "I've heard it before"; 4 = "I've played it before"). Scores averaged across listeners indicated low familiarity with the music (1.94 and 2.00 for Haydn movements I and II; 1.88 and 1.56 for Debussy movements I and II).

3.2. Materials and Equipment

Listeners heard the music from Creative A50 speakers, adjusted to a comfortable volume, while pupil data was collected from a stationary eye tracker (SMI iView RED) at 60 Hz. Listeners rested their chin and forehead on a chinrest positioned 70 cm from a computer screen. The experiment was run through SMI Experiment Center, which collected pupil data and presented audio/visual stimuli. During listening trials, the screen featured a white background with a black outline of a circle. Listeners were instructed to keep their eyes fixated within the circle.

3.3. Procedure

Participants listened to the recordings in the same order that they were performed (Haydn 1st movement, Haydn 2nd movement, Debussy 1st movement, Debussy 2nd movement). They were given the name and composer of each piece and asked to keep their eyes open and fixated on the computer screen while listening. A 60-s baseline pupil measurement was taken prior to each listening trial. Following each trial, the participants were asked to rate their familiarity with the piece they had just heard, and then allowed to take a break before continuing. Following the final listening trial, they answered some questions about their musical background.

3.4. Analysis

3.4.1. Preprocessing of Pupil Data

We used the same preprocessing procedure as in Experiment 1 (Section 2.4.1).

3.4.2. Linear Mixed Effects Modelling of Pupil Diameter

We used the same modelling procedure as in Experiment 1. For ratings of Technical difficulty, Harmonic complexity, and Expressive difficulty, we averaged the performers' ratings at each bar to get a single series of values per predictor. Similarly, for Head and Arm, we averaged quantity of motion values across performers at each bar. Effect sizes for significant predictors were estimated using the same hierarchically modelling procedure as in Experiment 1.

TABLE 5 | Results of linear mixed effects modelling for listeners.

Model	Fixed effect	Estimate	SE	*z*-value	χ^2	BIC
Model 1	(null model)	—	—	—	—	–3164.9
	Harmonic complexity	–0.0426	0.0040	10.60***	126.00***	–3281.7
	Expressive difficulty	0.0347	0.0041	8.52***	98.94***	–3371.4
	Technical difficulty	–0.0100	0.0002	2.50*	9.64**	–3371.9
	QoM head	–0.0023	0.0002	10.31***	103.47***	–3466.1
	Cloud diameter	–0.0008	0.0017	0.45	—	—
	Sound level	1.4e-5	9.3e-6	1.53	—	—
Model 2	(null model)	—	—	—	—	–3164.9
	Harmonic complexity	–0.0456	0.0040	11.36***	126.00***	–3281.7
	Expressive difficulty	0.0372	0.0041	9.13***	98.94***	–3371.4
	Technical difficulty	–0.0101	0.0040	2.51*	9.64**	–3371.9
	QoM arms	–0.0006	9.7e-5	6.20***	37.95***	–3400.6
	Cloud diameter	–9.7e-5	0.0017	0.06	—	—
	Sound level	6.8e-6	0.0040	0.73	—	—

*$p < 0.05$, **$p < 0.01$, *and* ***$p < 0.001$.

3.5. Results

The results of the LMM are given in **Table 5**. Both models showed a positive effect of Expressive difficulty and negative effects of Harmonic complexity and Technical difficulty on pupil diameter. Model 1 showed a negative effect of Head motion, and Model 2 showed a negative effect of Arm motion. Cloud diameter and Sound level did not yield significant effects for either model. The four significant predictors also showed significant χ^2 values when added to a null model and reduced the BIC (although the reduction by technical difficulty was very small). Descriptive plots for significant predictors are given in the **Supplementary Figures 1**–**3**,**5**,**6**.

3.6. Discussion

This experiment tested the effects of musical complexity, performers' bodily effort, and performers' ratings of expressive difficulty on pupil size in musically-trained listeners. In line with Experiment 1, we observed negative effects of musical complexity and a positive effect of expressive difficulty. In contrast to Experiment 1, this experiment showed significant negative effects of quantity of head and arm motion and performers' ratings of technical difficulty on listeners' pupil sizes. This result was not in line with our prediction that "traces" of performers' bodily effort would "sound" through the music and demand increased mental effort and arousal during listening.

In the literature, there is convergent evidence from many brain imaging and behavioural studies that music listening engages motor circuits (Novembre and Keller, 2014). However, questions remain regarding the extent to which this motor engagement is necessarily an active, attention-drawing process. We presented our participants with a passive listening task about 20 min in duration. Though all of our participants were trained musicians, their musical background and interests varied and several of them were not familiar with string quartet repertoire. Thus, some of our participants may not have been very actively engaged in listening. Furthermore, to maximize quality of pupil

data, we did not include video in the stimulus presentation. Visual presentation of performers' gestures can activate covert simulation mechanisms in observers, and might have encouraged motor engagement from our participants (Haslinger et al., 2005; Wöllner and Cañal-Bruland, 2010; Su and Pöppel, 2012; Taylor and Witt, 2014).

We observed a negative effect of performers' ratings of Harmonic complexity on pupil size. The effect of Cloud diameter was also negative, but nonsignificant. This finding is in line with the results of Experiment 1, and in conflict with our original prediction that harmonic complexity would have a stimulating effect on mental effort. We propose an explanation for these findings in the General Discussion.

Finally, we observed a positive effect of expressive difficulty for listeners. This is in line with our prediction, though in contrast to our findings in Experiment 1, where expressive difficulty showed no effect. The larger sample size in Experiment 2 may have allowed this effect to come through. Expressive difficulty or intensity may also account for more of the variance in pupil size during music listening than during music performance, since there are fewer demands on the motor system. We explore this effect in greater depth in the General Discussion.

4. GENERAL DISCUSSION

This study investigated the effects of harmonic complexity, bodily effort, and expressive difficulty on mental effort and arousal during music performance and listening. Pupil diameter was used to estimate mental effort and arousal. In Experiment 1, we collected pupil data from the members of a student string quartet as they performed in rehearsal and concert settings in our lab. In Experiment 2, we collected pupil data from a sample of trained musicians as they listened to the quartet's concert recordings.

Our results revealed stimulating effects of bodily effort and expressive difficulty on performers' pupil size, and stimulating effects of expressive difficulty on listeners' pupil size, in line with our predictions. Contrary to our predictions, we also observed consistently negative effects of Harmonic complexity and Cloud diameter (i.e., harmonic dissonance) for both performers and listeners, and negative effects of performers' bodily effort on listeners. Finally, we saw elevated levels of arousal in both violinists during their turns as 1st violinist. These findings are discussed in more detail below.

4.1. Body Motion, Technical Difficulty, and Sound Level

Performers' arm motion and perceived technical difficulty had stimulating effects on their pupil diameter. These effects supported our prediction that bodily effort would contribute to mental effort and arousal. Our results showed that arm motion improved the fit of a model that already included technical difficulty, suggesting that it accounted for a unique part of the variance. These findings are in line with previous studies of musical effort, which show greater pupil dilation in overt performance than in listening or imagined performance, suggesting an increased demand on cognitive resources (O'Shea and Moran, 2019; Endestad et al., 2020). A remaining question is whether we can identify unique effects of motor exertion and the mental effort involved in motor control on pupil size. The design of the current study did not enable us to make this distinction; however, future studies might present performers with a task that independently varies exertion and complexity.

Our other two measures of bodily effort—quantity of sound level and head motion—did not yield significant effects. The lack of effect for sound level suggests that this is not a strong predictor of mental effort, despite the relationship between sound level and physical force. Head motion, for string players, is primarily expressive and does not generally require as much motor control or physical strength as does arm motion, which is directly involved in sound production (see Discussion 1). This could partially explain the lack of stimulating effect of head motion on pupil size.

Expressive non-sound-producing head or body motion might indeed require relatively little mental effort overall, especially for experienced musicians who are playing well-practiced repertoire, and might even facilitate structuring of the performance. Expressive body motion is an integral component of expressive performance (Glowinski et al., 2013b; Chang et al., 2019). Skilled musicians reduce their body motion substantially when asked to play deadpan, even if no specific instructions regarding body motion are given (Davidson, 2007; Thompson and Luck, 2012). When asked to given an "immobile" performance, however, some slight expressive motion persists (Wanderley, 2002; Wanderley et al., 2005). Thus, to some degree, expressive body motion may occur automatically as a result of the performer's embodied relationship with the music (van der Schyff et al., 2018; Høffding and Satne, 2019). Still, further research is needed to show under what conditions expressive body motion requires more mental effort. In particular, it would be interesting to test whether expressive body motion reduces when other aspects of performing increase in difficulty.

For listeners, performers' head and arm motion and ratings of technical difficulty had a negative effect on pupil size. As we explained in the Discussion of Experiment 2, the lack of positive effect might be attributable to participants adopting a passive listening style during the experiment. Listeners might also not have perceived such substantial variability in the technical demands that were presented by the Haydn and Debussy selections, or they might have perceived difficulty across longer timeframes than we accounted for. For example, perhaps we would see different responses to music with long alternating periods of greater and lesser technical difficulty. Our measures of performers' bodily effort might also co-vary with another aspect of musical structure that our analysis did not capture (e.g., changes in timbre or tone quality, or phrase structure), that had a stronger effect on mental effort, resulting in a seemingly negative relationship between bodily effort and pupil size. While it might be useful in this case to consider other variables relating to musical structure as possible co-predictors, with the modelling approach that we used, care must be taken to avoid overfitting. A more effective approach might be to compare pupil responses to stimuli that vary more strongly in their physical and technical demands.

4.2. Harmonic Complexity, Tonal Tension, and Expressivity

Both experiments revealed negative effects of harmonic complexity (measured using performers' ratings and Cloud diameter) on pupil size, in contrast to our predictions. Harmonic complexity is just one component of musical complexity. While it might substantially engage listeners' attention when all other components are constant, this does not happen in real music; instead, different structural components might vary simultaneously, forming combined demands on attention that fluctuate over time. Thus, the negative effect that we saw on pupil size may reflect the relationships that harmonic complexity shares with other (untested) components of musical structure.

We should also note that for repertoire in the Western classical music tradition prior to the early twentieth century, harmonic complexity varies within fairly strict bounds (although these bounds changed over time, and differ between the Haydn and Debussy selections that we studied). In some music, periods of extreme harmonic complexity (e.g., as we see in some contemporary classical music that makes use of atonality) might make heightened demands on attention that outweigh the demands made by other structural features. In such cases, we might see a clear relationship between harmonic complexity and pupil size, despite the fact that harmony is embedded in a larger musical structure (and despite the passive rather than active/analytical listening task). However, even the more varied harmony present in Debussy's music is unlikely to achieve these extremes. During periods of high complexity, we might alternatively expect to see effects of listeners "zoning out" (mind wandering) if following the music proves too difficult or effortful (Unsworth and Robison, 2017; O'Shea and Moran, 2019). This would cause a reduction in pupil size. However, we would expect this reduction to endure over a relatively long period of time, not fluctuate from bar to bar.

Overall, we interpret these results as suggesting that individual low-level components of musical structure may not provide useful predictions of mental effort for performers or listeners. Higher-order predictors, which represent combinations of structural components, may be more effective. Indeed, performers' ratings of expressive difficulty—a measure which we presume is informed by musical complexity, technical difficulty, and emotional intensity—did predict pupil size for performers and listeners. The effect was slightly weaker for performers (significant only in Model 2), possibly because of conflicting demands by the motor system.

For ensemble players, an additional higher-order variable that might relate to pupil size is subjective ratings of coordination difficulty (how difficult is it for the ensemble to play together as a unit?). This variable would draw on factors relating to complexity, expressivity, and technical difficulty, as well as the relationships between parts in the ensemble (Keller et al., 2014)[1]. For classical ensembles playing well-practiced music, most of the major interpretive decisions have already been made, but coordination can still be challenging if the performers vary aspects of their practiced performance. Certain musical passages may continue to pose a challenge even after some practice; for example, long pauses followed by synchronized chords may continue to require effortful coordination (Bishop et al., 2019b). Future research could use pupillometry measures to investigate the relationship between coordination demands resulting from different structural contexts and mental effort.

The results of this study could be interpreted in terms of the adaptive-gain theory, which posits that exploitative and explorative control states underlie optimal performance on behavioural tasks (Aston-Jones and Cohen, 2011). These control states are mediated by the locus coeruleus-norepinephrine system. Exploitation is associated with phasic LC activity, intermediate pupil sizes, and increased responsivity to task-relevant stimuli, while exploration is associated with tonic LC activity, large pupil sizes, and facilitated processing of task-irrelevant stimuli or behaviour (Jepma and Nieuwenhuis, 2011). Musicians may switch between exploitative and explorative modes during performance as the musical demands change, resulting in fluctuations in pupil size.

4.3. Use of Mobile Eye Tracking in Music Performance Settings

This study is the first to use pupillometry to explore the relationships between cognitive-motor task demands and mental effort during ensemble performance in an ecological setting. Outside of a controlled lab environment, there are several factors that may add noise to pupil data. Three primary types of pupil response have been described in the literature: the pupil light response, which involves a pupil constriction in response to increased brightness; the pupil near response, which involves a pupil constriction when gaze shifts from a further-away object to a nearer object; and the psychosensory pupil response, which involves a pupil dilation in response to increased arousal or mental effort (Mathôt, 2018).

In our data collection, the performers unavoidably encountered changes in brightness and shifted their focus between nearer and further-away objects as their gaze moved between different parts of the visual scene. We attempted to minimize the effects of brightness by controlling the lighting of the performance space (i.e., using a constant, moderately bright level of room lighting instead of any stage lighting) and by covering the walls and windows with black curtains, to avoid any stray lights shining into the space and lighting contrasts between the white walls and dark floor. These controlled lighting conditions were also needed to minimize extraneous reflections for motion capture. The musicians played from scores, which allowed for some consistency in terms of brightness and distance in the visual display; however, they did not look exclusively at their scores throughout the performances. Ensemble musicians often spend some playing time watching their co-performers (Vandemoortele et al., 2018; Bishop et al., 2019a), and this was also the case in the current study (see results in[2]).

Of course, our attempts to control the performance space did have some effect on the musicians' experience, reducing the ecological validity of the performance. In particular, the musicians struggled with the dry acoustics of the space. In future studies of mental effort in music performance, we would recommend having musicians play from a score or fix their eyes

on an empty screen or stand, although we would also note that instructing performers to restrict their gaze to the score can have some unintended effects on how they behave. Our analysis of the quartet's body motion showed lower quantity of motion in two rehearsal performances where they could look only at the score than in the other "free gaze" conditions see[2].

The lack of control in Experiment 1 might have introduced noise into our pupil data. More concerning is the possibility of systematic effects. For example, the musicians might have looked away from the score during "easier" passages, and returned their gaze to the score during more "difficult" passages. In such cases, pupil constrictions (triggered by looking at the relatively bright and nearby score) might occur in association with increasing piece difficulty. Our results showed the opposite effect for technical and expressive difficulty, which stimulated pupil dilation rather than constriction. We did observe a negative relationship between pupil dilation and harmonic complexity measures; however, the same negative effect occurred for listeners in Experiment 2.

We did not attempt to compensate for effects of brightness or distance in our data analysis. We might have removed data segments where performers were not looking at the score, for example, but this would have reduced our dataset dramatically, and much more for some musicians than others. The musicians differed greatly in what percentage of performance time they spent looking at the score vs. at their co-performers. The cellist, in particular, sometimes spent as little as 40% of performance time looking at the score. Instead, we chose to complement our performance experiment with a listening experiment, which included a larger sample of remote participants, who completed their task in controlled conditions in the laboratory. With the listening experiment, we were able to confirm some of the results suggested by our performers' data as well as address some additional predictions.

As we showed in this study, overt body motion has an effect on pupil size. Pupillometry studies using music performance (or sports performance, etc.) paradigms have a unique opportunity to examine this relationship. In our case, we organized a recording set-up that would allow us to collect synchronized motion capture and eye tracking/pupillometry data. With motion capture data for performers heads and arms, we were able to show how these different types of movement have different effects. As we mentioned above, difficulties in distinguishing between the effects of body motion and the effects of other task demands on pupil size can also be problematic for researchers who want to measure these effects in isolation. In such cases, careful construction of an experimental paradigm that controls for different task demands would be needed.

Our ecological data capture gave us the unique opportunity to document a performance disruption: the 1st Violinist's missed page turn during the concert performance of the 1st movement of the Haydn String Quartet, which necessitated him to play the final page by memory. His pupils remained dilated throughout the rest of the piece, indicating heightened demands on mental effort. This type of performance disruption, along with the performer's natural reaction, would be hard to capture (or trigger) under more controlled, less ecological conditions. Our capture of this disruption shows how the violinist was able to successfully cope with the disruption by continuing to play from memory, without overt acknowledgment of the error. His individual coping response made the group resilient to greater disruption from the error (Glowinski et al., 2016).

4.4. Conclusions

This study shows an overlap in how performers and listeners attend to string quartet music from the Western classical repertoire. Both performers and listeners responded to changes in expressivity and musical structure, and performers' arousal levels were predicted by their sound-producing arm motion. The violinists in the quartet also both showed heightened arousal when performing as first violinist, suggesting that heightened stress may be associated with this role. Our study also demonstrates how mental effort and arousal can be successfully assessed using eye tracking and motion capture technologies in a relatively naturalistic concert setting. In the future, it would be valuable to build on our findings with studies of how coordination difficulty contributes to mental effort in ensemble performance, and to distinguish between the effects of overt motion and the mental effort associated with motor control on performers' pupil size.

DATA AVAILABILITY STATEMENT

The data for this study are openly available in the Quartet Body Motion and Pupillometry Database (doi: 10.5281/zenodo.4888176).

ETHICS STATEMENT

This study was approved by the Norwegian Center for Research Data (NSD), with the project identification number 748915. Written informed consent was obtained from participants for the publication of potentially identifiable images and data.

AUTHOR CONTRIBUTIONS

LB, BL, and AJ conceived the experiment. LB collected the data, ran the analysis, and wrote the first draft of the paper. BL and AJ consulted on the analysis and contributed to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the University of Oslo and the Research Council of Norway through its Centres of Excellence scheme, Project Number 262762.

ACKNOWLEDGMENTS

We are grateful to Prof. Werner Goebl and Dr. Tetsuto Minami for the loan of their eye-tracking glasses. We wish to thank the Borealis String Quartet for their enthusiastic participation in this project. Finally, big thanks go to Dr. Victor Gonzalez Sanchez

for his work on the data collection, and to Assoc. Prof. Simon Høffding for initiating and organizing the collaboration with the string quartet and facilitating the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.653021/full#supplementary-material

Supplementary Figure 1 | Violin plots showing the distribution of Quantity of Arm Motion values for each musician across pieces.

Supplementary Figure 2 | Violin plots showing the distribution of ratings of Technical difficulty given by each musician across pieces.

Supplementary Figure 3 | Violin plots showing the distribution of ratings of Harmonic Complexity given by each musician across pieces.

Supplementary Figure 4 | Violin plots showing the distribution of Cloud diameter values across pieces.

Supplementary Figure 5 | Violin plots showing the distribution of ratings of Expressive difficulty given by each musician across pieces.

Supplementary Figure 6 | Violin plots showing the distribution of Quantity of Head Motion values for each musician across pieces.

REFERENCES

Abrams, D. A., Ryali, S., Chen, T., Chordia, P., Khouzam, A., Levitin, D. J., et al. (2013). Inter-subject synchronization of brain responses during natural music listening. *Eur. J. Neurosci.* 37, 1458–1469. doi: 10.1111/ejn.12173

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., and Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *J. Vis.* 14, 1–20. doi: 10.1167/14.4.1

Aston-Jones, G., and Cohen, J. D. (2011). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J. Compar. Neurol.* 493, 99–110. doi: 10.1002/cne.20723

Bianco, R., Ptaszynski, L. E., and Omigie, D. (2020). Pupil responses to pitch deviants reflect predictability of melodic sequences. *Brain Cogn.* 138:103621. doi: 10.1016/j.bandc.2019.103621

Bishop, L., Cancino-Chacón, C. E., and Goebl, W. (2019a). Eye gaze as a means of giving and seeking information during musical interaction. *Consciousness Cogn.* 68, 73–96. doi: 10.1016/j.concog.2019.01.002

Bishop, L., Cancino-Chacón, C. E., and Goebl, W. (2019b). Moving to communicate, moving to interact: Patterns of body motion in musical duo performance. *Music Percept.* 37, 1–25. doi: 10.1525/mp.2019.37.1.1

Bishop, L., and Goebl, W. (2017). "Music and movement: musical instruments and performers," in *The Routledge Companion to Music Cognition*, eds R. Ashley and R. Timmers (New York, NY: Routledge), 349–361.

Bishop, L., and Goebl, W. (2018). Beating time: how ensemble musicians' cueing gestures communicate beat position and tempo. *Psychol. Music* 46, 84–106. doi: 10.1177/0305735617702971

Bishop, L., and Jensenius, A. R. (2020). "Reliability of two IR motion capture systems in a music performance setting," in *Proceedings of the International Conference on Sound and Music Computing* (Torino).

Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., and Widmer, G. (2016). "madmom: a new Python Audio and Music Signal Processing Library," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, 1174–1178.

Bojner Horwitz, E., Harmat, L., Osika, W., and Theorell, T. (2020). The interplay between chamber musicians during two public performances of the same piece?a novel methodology using the concept of 'flow'. *Front. Psychol.* 11:618227. doi: 10.3389/fpsyg.2020.618227

Burke, M. J., and Gridley, M. C. (1990). Musical preferences as a function of stimulus complexity and listeners' sophistication. *Percept. Motor Skills* 71, 687–690. doi: 10.2466/PMS.71.6.687-690

Campbell, M. J., Moran, A. P., Bargary, N., Surmon, S., Bressan, L., and Kenny, I. C. (2019). Pupillometry during golf putting: a new window on the cognitive mechanisms underlying quiet eye. *Sport Exerc. Perform. Psychol.* 8, 53–62. doi: 10.1037/spy0000148

Cancino-Chacón, C. E., Gadermaier, T., Widmer, G., and Grachten, M. (2017). An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music. *Mach. Learn.* 106, 887–909. doi: 10.1007/s10994-017-5631-y

Chaffin, R., Lisboa, T., Logan, T., and Begosh, K. T. (2010). Preparing for memorized cello performance: the role of performance cues. *Psycho. Music* 38, 3–30. doi: 10.1177/0305735608100377

Chaffin, R., and Logan, T. (2006). Practicing perfection: how concert soloists prepare for performance. *Adv. Cogn. Psychol.* 2, 113–130. doi: 10.2478/v10053-008-0050-z

Chang, A., Kragness, H., Livingstone, S., Boxnyak, D. J., and Trainor, L. J. (2019). Body sway reflects joint emotional expression in music ensemble performance. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-018-36358-4

Chew, E. (2000). *Towards a Mathematical Model of Tonality*. Ph.D. thesis, Massachusetts Institute of Technology.

Dalmazzo, D., and Ramírez, R. (2019). Bowing gestures classification in violin performance: a machine learning approach. *Front. Psychol.* 10, 1:344. doi: 10.3389/fpsyg.2019.00344

D'Amato, V., Volta, E., Oneto, L., Volpe, G., Camurri, A., and Anguita, D. (2020). Understanding violin players' skill level based on motion capture: a data-driven perspective. *Cogn. Comput.* 12, 1356–1369. doi: 10.1007/s12559-020-09768-8

Davidson, J. W. (2007). Qualitative insights into the use of expressive body movement in solo piano performance: a case study approach. *Psychol. Music* 35, 381–401. doi: 10.1177/0305735607072652

Davidson, J. W., and Good, J. M. (2002). Social and musical co-ordination between members of a string quartet: an exploratory study. *Psychol. Music* 30, 186–201. doi: 10.1177/0305735602302005

Egermann, H., Fernando, N., Chuen, L., and McAdams, S. (2015). Music induces universal emotion-related psychophysiological responses: comparing Canadian listeners to Congolese Pygmies. *Front. Psychol.* 5:1341. doi: 10.3389/fpsyg.2014.01341

Egermann, H., Pearce, M. T., Wiggins, G. A., and McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cogn. Affect. Behav. Neurosci.* 13, 533–553. doi: 10.3758/s13415-013-0161-y

Eitan, Z., and Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition* 114, 405–422. doi: 10.1016/j.cognition.2009.10.013

Endestad, T., Godøy, R. I., Sneve, M. H., Hagen, T., Bochynska, A., and Laeng, B. (2020). Mental effort when playing, listening, and imagining music in one Pianist's eyes and brain. *Front. Hum. Neurosci.* 14:576888. doi: 10.3389/fnhum.2020.576888

Fink, L. K., Hurley, B. K., Geng, J. J., and Janata, P. (2018). A linear oscillator model predicts dynamic temporal attention and pupillary entrainment to rhythmic patterns. *J. Eye Mov. Res.* 11, 1–25. doi: 10.16910/jemr.11.2.12

Gingras, B., Marin, M. M., Puig-Waldmüller, E., and Fitch, W. T. (2015). The eye is listening: music-induced arousal and individual differences predict pupillary responses. *Front. Hum. Neurosci.* 9, 1–12. doi: 10.3389/fnhum.2015.00619

Glowinski, D., Bracco, F., Chiorri, C., and Grandjean, D. (2016). Music ensemble as a resilient system. managing the unexpected through group interaction. *Front. Psychol.* 7:1548. doi: 10.3389/fpsyg.2016.01548

Glowinski, D., Dardard, F., Gnecco, G., Piana, S., and Camurri, A. (2015). Expressive non-verbal interaction in a string quartet: an analysis through head movements. *J. Multimodal User Interfaces* 9, 55–68. doi: 10.1007/s12193-014-0154-3

Glowinski, D., Gnecco, G., Piana, S., and Camurri, A. (2013a). "Expressive non-verbal interaction in string quartet," in *Proceedings-2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (Geneva), 233–238.

Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., and Doherty, C. (2013b). The movements made by performers in a skilled quartet: a distinctive pattern, and the function that it serves. *Front. Psychol.* 4:841. doi: 10.3389/fpsyg.2013.00841

Godøy, R. I., Haga, E., and Jensenius, A. R. (2005). "Playing "air instruments": mimicry of sound-producing gestures by novices and experts," in *Gesture in Human-Computer Interaction and Simulation*, Berlin: Springer-Verlag, 256–267.

Gordon, C. L., Cobb, P. R., and Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: an ALE meta-analysis of fMRI data. *PLoS ONE* 13:e0207213. doi: 10.1371/journal.pone.0207213

Gordon, J., and Gridley, M. C. (2013). Musical preferences as a function of stimulus complexity of piano jazz. *Creat. Res. J.* 25, 143–146. doi: 10.1080/10400419.2013.752303

Grachten, M., Cancino-Chacón, C., and Gadermaier, T. (2019). "partitura: a python package for handling symbolic musical data," in *Late Breaking/Demo at the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, Delft.

Haslinger, B., Erhard, P., Altenmüller, E., Schroeder, U., Boecker, H., and Ceballos-Baumann, A. O. (2005). Transmodal sensorimotor networks during action Observation in professional pianists. *J. Cogn. Neurosci.* 17, 282–293. doi: 10.1162/0898929053124893

Haueisen, J., and Knösche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. *J. Cogn. Neurosci.* 13, 786–792. doi: 10.1162/08989290152541449

Herremans, D., and Chew, E. (2016). "Tension ribbons: quantifying and visualising tonal tension," in *Proceedings of TENOR* (Cambridge), 1–10.

Høffding, S. and Satne, G. (2019). Interactive expertise in solo and joint musical performance. *Synthese* 198, 427–445. doi: 10.1007/s11229-019-02339-x

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

Jensenius, A. R., Wanderley, M. M., Godoy, R. I., and Leman, M. (2010). "Musical gestures: Concepts and methods in research," in *Musical Gestures: Sound,Movement, and Meaning*, eds R. I. Godoy and M. Leman (New York, NY: Routledge), 12–35.

Jepma, M., and Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. *J. Cogn. Neurosci.* 23, 1587–1596. doi: 10.1162/jocn.2010.21548

Joshi, S., Li, Y., Kalwani, R. M., and Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* 89, 221–234. doi: 10.1016/j.neuron.2015.11.028

Juslin, P. N. (2003). Five facets of musical expression: a psychologist's perspective on music performance. *Psychol. Music* 31, 273–302. doi: 10.1177/03057356030313003

Just, M. A., Carpenter, P. A., and Miyake, A. (2003). Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. *Theor. Issues Ergonomics Sci.* 4, 59–88. doi: 10.1080/14639220210159735

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.

Kahneman, D., and Beatty, J. (1966). Pupil iameter and load on memory. *Science* 154, 1583–1585. doi: 10.1126/science.154.3756.1583

Kang, O., and Banaji, M. R. (2020). Pupillometric decoding of high-level musical imagery. *Consciousness Cogn.* 77:102862. doi: 10.1016/j.concog.2019.102862

Kang, O., and Wheatley, T. (2015). Pupil dilation patterns reflect the contents of consciousness. *Conscious. Cogn.* 35, 128–135. doi: 10.1016/j.concog.2015.05.001

Kelkar, T., and Jensenius, A. R. (2018). Analyzing free-hand sound-tracings of melodic phrases. *Appl. Sci.* 8:135. doi; 10.3390/app8010135

Keller, P. E. (2001). Attentional resource allocation in musical ensemble performance. *Psychol. Music* 29, 20–38. doi: 10.1177/0305735601291003

Keller, P. E., Novembre, G., and Hove, M. J. (2014). Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130394. doi: 10.1098/rstb.2013.0394

Kenny, D. T. (2011). *The Psychology of Music Performance Anxiety*. Oxford: Oxford University.

Klingner, J., Tversky, B., and Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology* 48, 323–332. doi: 10.1111/j.1469-8986.2010.01069.x

Konecni, V. J. (2005). The aesthetic trinity: awe, being moved, thrills. *Bull. Psychol. Arts* 5, 27–44. doi: 10.1037/e674862010-005

Konishi, M., Brown, K., Battaglini, L., and Smallwood, J. (2017). When attention wanders: Pupillometric signatures of fluctuations in external attention. *Cognition* 168, 16–26. doi: 10.1016/j.cognition.2017.06.006

Laeng, B., and Alnæs, D. (2019). "Pupillometry," in *Eye Movement Research*, eds C. Klein and U. Ettinger (Switzerland, Springer Nature), 449–502.

Laeng, B., Eidet, L. M., Sulutvedt, U., and Panksepp, J. (2016). Music chills: the eye pupil as a mirror to music's soul. *Consciousness Cogn.* 44, 161–178. doi: 10.1016/j.concog.2016.07.009

Laeng, B., Ørbo, M., Holmlund, T., and Miozzo, M. (2011). Pupillary stroop effects. *Cogn. Proc.* 12, 13–21. doi: 10.1007/s10339-010-0370-z

Lenartowicz, A., Simpson, G. V., and Cohen, M. S. (2013). Perspective: causes and functional significance of temporal variations in attention control. *Front. Hum. Neurosci.* 7, 1–7. doi: 10.3389/fnhum.2013.00381

Liao, H. I., Yoneya, M., Kidani, S., Kashino, M., and Furukawa, S. (2016). Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention. *Front. Neurosci.* 10:43. doi: 10.3389/fnins.2016.00043

Lundqvist, L. O., Carlsson, F., Hilmersson, P., and Juslin, P. N. (2009). Emotional responses to music: experience, expression, and physiology. *Psychol. Music* 37, 61–90. doi: 10.1177/0305735607086048

Maes, P. J., Leman, M., Palmer, C., and Wanderley, M. M. (2014). Action-based effects on music perception. *Front. Psychol.* 4:1008. doi: 10.3389/fpsyg.2013.01008

Marin, M. M., and Leder, H. (2013). Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PLoS ONE* 8:e72412. doi: 10.1371/journal.pone.0072412

Marois, A., Labonté, K., Parent, M., and Vachon, F. (2018). Eyes have ears: indexing the orienting response to sound using pupillometry. *Int. J. Psychophysiol.* 123, 152–162. doi: 10.1016/j.ijpsycho.2017.09.016

Mathôt, S. (2018). Pupillometry: psychology, physiology, and function. *J. Cogn.* 1, 1–23. doi: 10.5334/joc.18

Mauch, M., and Levy, M. (2011). "Structural change on multiple time scales as a correlate of musical complexity," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, 489–494.

Menninghaus, W., Wagner, V., Hanich, J., Wassiliwizky, E., Kuehnast, M., and Jacobsen, T. (2015). Towards a psychological construct of being moved. *PLoS ONE* 10:e0128451. doi: 10.1371/journal.pone.0128451

Molnar-Szakacs, I., Green Assuied, V., and Overy, K. (2011). "Shared affective motion experience (SAME) and creative, interactive music therapy," in *Musical Imaginations: Multidisciplinary Perspectives on Creativity, Performance, and Perception*, eds D. J. Hargreaves, D. Miell, and R. MacDonald (Oxford: Oxford University), 313–331.

Murphy, P. R., Van Moort, M. L., and Nieuwenhuis, S. (2016). The pupillary orienting response predicts adaptive behavioral adjustment after errors. *PLoS ONE* 11:e0151763. doi: 10.1371/journal.pone.0151763

Novembre, G., and Keller, P. E. (2014). A conceptual review on action-perception coupling in the musicians' brain: what is it good for? *Front. Hum. Neurosci.* 8:603. doi: 10.3389/fnhum.2014.00603

Olsen, K. N., and Dean, R. T. (2016). Does perceived exertion influence perceived affect in response to music? Investigating the "FEELA"9D hypothesis. *Psychomusicology* 26, 257–269. doi: 10.1037/pmu0000140

O'Shea, H., and Moran, A. (2019). Are fast complex movements unimaginable? Pupillometric studies of motor imagery in expert piano playing. *J. Motor Behav.* 51, 371–384. doi: 10.1080/00222895.2018.1485010

Palmer, C. (1997). Music performance. *Annu. Rev. Psychol.* 48, 115–153. doi: 10.1146/annurev.psych.48.1.115

Papageorgi, I., Hallam, S., and Welch, G. F. (2007). A conceptual framework for understanding musical performance anxiety. *Res. Stud. Music Edu.* 28, 83–107.

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Ann. N. Y. Acad. Sci.* 1423, 378–395. doi: 10.1111/nyas.13654

Peifer, C., Schulz, A., Schächinger, H., Baumann, N., and Antoni, C. H. (2014). The relation of flow-experience and physiological arousal under stress-Can u shape it? *J. Exp. Soc. Psychol.* 53, 62–69. doi: 10.1016/j.jesp.2014.01.009

Piras, A., Timmis, M., Trofè, A., and Raffi, M. (2020). Understanding the underlying mechanisms of Quiet Eye: The role of microsaccades, small saccades and pupil-size before final movement initiation in a soccer penalty kick. *Eur. J. Sport Sci.* 15, 1–10. doi: 10.1080/17461391.2020.1788648

Repp, B. H., and Knoblich, G. (2007). Action can affect auditory perception. *Psychol. Sci.* 18, 6–7. doi: 10.1111/j.1467-9280.2007.01839.x

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.* 10, 211–223. doi: 10.1038/nrn2573

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Percept.* 21, 561–585. doi: 10.1525/mp.2004.21.4.561

Skaansar, J. F., Laeng, B., and Danielsen, A. (2019). Microtiming and mental effort: Onset asynchronies in musical rhythm modulate pupil size. *Music Percept.* 37, 111–133. doi: 10.1525/mp.2019.37.2.111

Su, Y. H., and Pöppel, E. (2012). Body movement enhances the extraction of temporal structures in auditory sequences. *Psychol. Res.* 76, 373–382. doi: 10.1007/s00426-011-0346-3

Taylor, J. E. T., and Witt, J. K. (2014). Listening to music primes space: pianists, but not novices, simulate heard actions. *Psychol. Res.* 79, 175–182. doi: 10.1007/s00426-014-0544-x

Thompson, M. R., and Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Sci.* 16, 19–40. doi: 10.1177/1029864911423457

Timmers, R., Endo, S., Bradbury, A., and Wing, A. M. (2014). Synchronization and leadership in string quartet performance: a case study of auditory and visual cues. *Front. Psychol.* 5:645. doi: 10.3389/fpsyg.2014.00645

Unsworth, N., and Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: a latent variable pupillometry study. *J. Exp. Psychol.* 43, 1962–1987. doi: 10.1037/xlm0000421

van der Schyff, D., Schiavio, A., Walton, A. E., Velardo, V., and Chemero, A. (2018). Musical creativity and the embodied mind : exploring the possibilities of 4E cognition and dynamical systems theory. *Music Sci.* 1, 1–18. doi: 10.1177/2059204318792319

van der Wel, P., and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: a review. *Psych. Bull. Rev.* 25, 2005–2015. doi: 10.3758/s13423-018-1432-y

Vandemoortele, S., Feyaerts, K., Reybrouck, M., De Bièvre, G., Brône, G., and De Baets, T. (2018). Gazing at the partner in musical trios: A mobile eye-tracking study. *J. Eye Mov. Res.* 11, 1–13. doi: 10.16910/jemr.11.2.6

Vickers, J. N. (2009). Advances in coupling perception and action: the quiet eye as a bidirectional link between gaze, attention, and action. *Progr. Brain Res.* 174, 279–288. doi: 10.1016/S0079-6123(09)01322-3

Vroegh, T. (2019). Zoning-in or tuning-in? Identifying distinct absorption states in response to music. *Psychomusicology* 29, 156–170. doi: 10.1037/pmu0000241

Vuvan, D. T., Simon, E., Baker, D. J., Monzingo, E., and Elliott, E. M. (2020). Musical training mediates the relation between working memory capacity and preference for musical complexity. *Memory Cogn.* 48, 972–981. doi: 10.3758/s13421-020-01031-7

Wanderley, M. M. (2002). "Quantitative analysis of non-obvious performer gestures," in *Gesture and Sign Language in Human-Computer Interaction*, eds I. Wachsmuth and T. Sowa (Berlin: Springer Verlag), 241–478.

Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., and Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *J. New Music Res.* 34, 97–113. doi: 10.1080/09298210500124208

Weiss, M. W., Trehub, S. E., Schellenberg, E. G., and Habashi, P. (2016). Pupils dilate for vocal or familiar music. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 1061–1065. doi: 10.1037/xhp0000226

Wilson, M., and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* 131, 460–473. doi: 10.1037/0033-2909.131.3.460

Wöllner, C., and Cañal-Bruland, R. (2010). Keeping an eye on the violinist: motor experts show superior timing consistency in a visual perception task. *Psychol. Res.* 74, 579–585. doi: 10.1007/s00426-010-0280-9

Yoshie, M., Kudo, K., Murakoshi, T., and Ohtsuki, T. (2009). Music performance anxiety in skilled pianists: Effects of social-evaluative performance situation on subjective, autonomic, and electromyographic reactions. *Exp. Brain Res.* 199, 117–126. doi: 10.1007/s00221-009-1979-y

Zekveld, A. A., Koelewijn, T., and Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: current state of knowledge. *Trends Hear.* 22, 1–25. doi: 10.1177/2331216518777174

Zénon, A., Sidibé, M., and Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Front. Behav. Neurosci.* 8:286. doi: 10.3389/fnbeh.2014.00286

Zickfeld, J. H., Schubert, T. W., Seibt, B., Blomster, J. K., Arriaga, P., Basabe, N., et al. (2019). Kama muta: conceptualizing and measuring the experience often labelled being moved across 19 nations and 15 languages. *Emotion* 19, 402–424. doi: 10.1037/emo0000450

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fpsyg.2021.674511

Adaptive Gaze Behavior and Decision Making of Penalty Corner Strikers in Field Hockey

Stefanie Klatt[1,2,3]*, *Benjamin Noël*[2], *Alessa Schwarting*[2], *Lukas Heckmann*[2] and *Frowin Fasold*[2]

[1] Institute of Sports Science, University of Rostock, Rostock, Germany, [2] Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Cologne, Germany, [3] School of Sport and Health Sciences, University of Brighton, Eastbourne, United Kingdom

In recent years, studies have increasingly dealt with the interaction of gaze behavior and decision making of team sports athletes. However, there is still a variety of important game situations, for example, in the case of penalty corners in field hockey, in which this interaction has not been investigated in detail yet. Penalty corners present a meaningful goal scoring opportunity by providing a relatively free shot. This paper considers two studies. The first study investigated a possible connection between the gaze behavior and the quality of decisions of experienced field hockey players and evaluated the level of success of different gaze strategies. A preliminary study (Study 1) was designed as a survey questionnaire with the aim of preparing for the main study by obtaining subjective assessments of the individual gaze behavior and decision making of professional athletes. In the second and the main study (Study 2), the gaze behavior of experienced field hockey players was recorded using mobile eye-tracking systems to analyze different strategical approaches in associated gaze behavior and decision making. Study 1 showed that players consider reacting to the defenders' behavior during a penalty corner a promising avenue for improving success at penalty corner attempts. It also indicated that such defense-dependent strategies are currently only rarely employed. Study 2 demonstrated how gaze behavior differs between different strategical approaches of the offense. It was shown that the gaze direction on the ball, the stopper, and the goal area is important to allow for a more optimal adaptation to the tactical behavior of defense. It can be concluded that adaptive decision making (i.e., choosing which variation will be carried out just after the "injection" of the ball) seems promising but requires further training to improve the success rate of penalty corner.

Keywords: drag-flick, eye-tracking, performance, sport expertise, tactical decision

OPEN ACCESS

Edited by:
Russell A. Cohen Hoffing, CCDC Army Research Laboratory, Human Research and Engineering, US Army Research Laboratory, United States

Reviewed by:
Christian Vater, University of Bern, Switzerland
Pradipta Biswas, Indian Institute of Science (IISc), India

***Correspondence:**
Stefanie Klatt
stefanie.klatt@uni-rostock.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *01 March 2021*
Accepted: *05 July 2021*
Published: *02 August 2021*

Citation:
Klatt S, Noël B, Schwarting A, Heckmann L and Fasold F (2021) Adaptive Gaze Behavior and Decision Making of Penalty Corner Strikers in Field Hockey. Front. Psychol. 12:674511. doi: 10.3389/fpsyg.2021.674511

Nowadays, a lot of games are decided by penalty corners in field hockey. For example, during the 2018 World Cup, the four group winners scored 38% of their goals through penalty corners in the preliminary rounds, while about a third of the goals scored against the four teams eliminated in the preliminary rounds were scored through penalty corners. Surprisingly however, there is currently little research on arguably the most important action in field hockey (for a study on indoor field hockey see Vinson et al., 2013). Only recently some studies have mainly addressed the role of the goalkeeper in defending penalty corners (Morris-Binelli et al., 2020, 2021). Our study addresses this game situation and tries to scrutinize if more (online) adaptation on offense can be considered a promising future development of behavior during penalty corner by questioning (Study 1) and testing (Study 2) expert field hockey players.

During a field hockey game, the penalty corner (or short corner) is considered a highly complex strategic part (Laird and Sutherland, 2003). It is awarded for fouls committed by the defending team within the goal-scoring circle or for serious fouls outside of this circle (intentional rule violations). Before the penalty corner is carried out, four defending players and the goalkeeper position themselves behind the backline, either next to or inside the goal. All other defending players must be beyond the centerline. Any number of players of the attacking team can position themselves outside the circle with sticks, hand, and feet not touching the ground inside the circle. One player from the attacking team, the "injector," passes the ball from the backline at least 10 meters from the goalpost, usually to a player stopping the ball (stopper) for another player (striker) who either shoots the ball directly or passes it to yet another player (who may or may not take a shot). After the ball has been played by the injector, the defending players and the goalkeeper are allowed to enter the circle. The stopper of the attacking team must receive the ball outside of the shooting circle or else, briefly leave the circle after receiving the ball before any other actions can be taken.

Even though it seems that straight shots at the goal by the striker promise the greatest rate of success, there will possibly be a much greater demand for various penalty corner variations that demand online strategical decision making in the next few years due to adaptations in defenders' behaviors and lower success rates (average conversion rate World Cup 2018 $\approx$ 20.5%, Rowe, 2019). Defensive behavior and play calling are now increasingly taking into account that on offense in most cases a straight shot is called. Various analyses of penalty corners during past international tournaments (World Cups and Olympic Games) have shown that the penalty corner effectiveness depends on the fit between strategical approaches of the offense and defense, that is, whether a direct shot or a pass is performed (cf. Vizcaya, 2015) and how defensive players run-out to prevent goal scoring. This implies that the strikers could have an advantage if they were able to recognize the defensive behavior of the opponent's team early in order to choose a more promising option for the penalty corner.

This is also done in comparable situations in other team sports as the penalty kick in soccer. In this situation, the penalty taker can choose to decide on kick direction prior to running-up the ball and stay with that decision regardless of what the goalkeeper subsequently does. This strategy is termed keeper-independent strategy (Kuhn, 1988; van der Kamp, 2006). However, strikers can also choose to employ a keeper-dependent strategy in which they decide for a temporal target but mainly wait for the goalkeeper to commit to one side only to kick to the opposite side of the goal (Kuhn, 1988; van der Kamp, 2006; Noël et al., 2014). Importantly, both strategies are associated with fundamentally different patterns of gaze behavior probably reflecting the fact that penalty takers employing a keeper-dependent strategy rely on information on the goalkeeper' behavior (Noël and van Der Kamp, 2012). Furthermore, it was shown recently that the likelihood of scoring depends on the combination of the goalkeepers' behavior and the strategic approach of penalty takers (Noël et al., 2021).

In field hockey, an approach similar to the keeper-dependent strategy would create the opportunity to adapt to the run-up behavior of the defensive players after the ball is in play in order to choose either a straight shot or another penalty corner variation (even though, in penalty corners, it is not only the striker who has to adapt, but also his/her teammates). In this regard, it is important to consider which general variations are usually played and with which defensive tactics (see also Study 1). Therefore, the following offensive and defensive tactics are basic schemes that are subject to a high degree of variability. Due to the different possible variants, the positions and routes of the players differ according to the situation and rarely follow an exactly identical structure.

In international field hockey, the direct shot at goal is the variation that is most often played. By using the drag-flick, where the ball is dragged on the shaft of the hockey stick and flung at the goal, the striker can reach high velocity of the ball (Baker et al., 2009). Furthermore, the drag-flick allows to raise the ball over 460 mm (backboard height; Ibrahim et al., 2017), which is the critical height for hits (striking by using a swinging movement of the stick toward the ball) in penalty corners. Another option that is regularly chosen is the 90° variation. Here, the striker hooks the ball into the stick head during the initial movement of the flick. Instead of releasing the ball in front of the body in the direction of the goal, the striker rotates 90° to the left to a teammate who receives the ball and shoots it at the goal. Thereby the 90° variation is an offensive tactic allowing a different striker to shoot from a closer position to the goal with slightly more time to do so. This variation belongs to the general category of "pass to the left." For the deflection variation, the ball is brought into play by the injector, while an attacker runs into the zone between the penalty spot and the backline and lays down his/her hockey stick on the ground to deflect the ball. The striker takes a low shot at the goal using a drag-flick or a hit. The hockey stick that has been laid down by the other attacker deflects the ball either high or into another direction or both, so that the goalkeeper and the defensive players at the goal-line only have a very short time to react.

In general, two different defensive tactics are commonly chosen by the defense, both of which are initiated from the same starting formation in order not to provide the opponents with any cues for their decisions before the ball is in play. In the commonly used starting formation, three defenders stand to the right and one defender stands to the left of the goalkeeper (from the perspective of the offensive players looking at the goal)[1]. This formation is referred to as 3:1. Two options arise from this situation: *3:1 3:1 and 3:1 2:2.* In the 3:1 3:1 situation, one defender (first runner) of the block of three runs up to the striker in order to cover the right corner of the goal as soon as the ball is brought into play by the injector. Another one of the players (trailer) on this side remains positioned slightly offset behind in order to defend possible deflection variations. The remaining player on this side stays on the goal line to be able to parry the shots that cannot be prevented by the first defender. The defender to the left of the goalkeeper positions himself/herself between the penalty spot and the goal-line to be able to clear possible rebounds from

[1]For depictions of the tactical approaches on offense and defense see **Supplementary Figures 1–4**.

the zone in front of the goalkeeper. Due to the fact that the right corner of the goal is defended directly by two players, the goalkeeper has the possibility of taking a step toward the left goalpost and of focusing more strongly on this corner of the goal. In the end position, three defenders stand to the right and one defender stands to the left of the goalkeeper (3:1).

In the *3:1 2:2 situation,* from the starting formation, the defender standing to the left of the goalkeeper runs toward the pass to the left of the striker. One defender of the block of three to the right of the goalkeeper runs up to the striker to cover the right corner of the goal. Another one of the players moves behind the goalkeeper to the left side and positions himself/herself in the zone between the penalty spot and the goal line. His/her task is to clear possible rebounds from the zone in front of the goalkeeper and defend variations. The remaining defender on the right side stays on the goal line in order to parry shots that cannot be prevented by the first defender. In the end position, two defenders are standing to the left and two to the right of the goalkeeper (2:2).

The two defense variations each prevent another possible attack strategy, respectively. By using the 3:1 3:1 the defenders are positioned closely around the penalty spot and supposed to defend possible deflection strategies. The 3:1 2:2 is supposed to defend the pass to the left (here 90° variation). Regardless of the defense variation, the drag-flick (direct shot) always represents a reasonable variant and an effective shooting technique when it comes to the penalty corner (Piñeiro et al., 2007; Rosalie et al., 2017), although, it cannot always be considered optimal because of the defenders' positioning at the goal-line. For the defense variation 3:1 the 90° variation represents an optimal counter strategy, because no defender is directly assigned to the attacker who receives the ball from the striker. Whereas, for the 2:2 variation, it represents a less appropriate counter strategy because in this case the player on the 90° position is directly defended. In contrast, the deflection variation is the optimal solution for the 2:2 variation, because of the relatively wide-open zone between the two defenders who are defending the attackers at the top of the circle and the two defenders focusing more on the zone close to the goal. While the defense is using a 3:1 variation a deflection variation seems ineffective in comparison to the 90° variation referring to the higher number of defenders around the penalty spot possibly interfere the execution of the variation.

However, in reacting on the defensive players' run-out, strikers would probably also have to deal with the same problems as penalty takers in soccer employing a keeper-dependent strategy. On the hand, they have to focus on the behavior of the defensive players but on the other hand they also have to prepare for/focus on the execution of their own actions (Noël and van Der Kamp, 2012). However, currently it is not known if and how often offensive players try to employ a "defense-dependent" strategy, but it appears that in the majority of the cases a "defense-independent" strategy is employed. It is also not known if it is possible to focus on aspects of defenders' behaviors while also preparing and coordinating self-actions given that it takes <2,000 ms from the ball being injected till the ball leaves the strikers stick. Furthermore, in penalty corners it is not only the striker who has to come to a reasonable decision in time, but also the teammates have to perceive the situation correctly to act accordingly. It is certainly possible that play callers consider these demands on the gaze behavior of players a major problem and therefore rely on a defense-independent strategy only in which every player on offense knows prior to the injection of the ball which variant (e.g., a direct shot) will be played.

Although, in general, a number of investigations of the combined gaze and decision-making behavior in high-performance sports is available (for reviews, see Kredel et al., 2017; Hüttermann et al., 2018), focusing on predominantly foveal (e.g., Noël and van Der Kamp, 2012) as well as peripheral vision (Vater et al., 2017), in field hockey, only Roth et al. (2007) have evaluated the gaze behavior of strikers so far. It was found that, as the ball is put into play, the defensive players running out of the circle had their gaze fixated by the striker in order for them to choose their action accordingly. As soon as the ball was stopped, only the ball was fixated by the striker. Thus, anticipatory processes may play a role in action selection since the strikers subsequently did not fixate on the space or the defenders anymore. Alternatively, peripheral vision can potentially be used by strikers through this process. Here, strikers may predominantly focus on the ball while monitoring other important aspects, like the stopper, peripherally (cf. Klatt and Smeeton, 2021a,b). In this way, such patterns of gaze may function as a gaze anchor (cf. Vater et al., 2016). However, because only two athletes were tested within the scope of these investigations by Roth et al. (2007) and the quality of the decision-making behavior was not considered, the meaningful relationship between gaze behavior and decision-making behavior remains unclear. Particularly this aspect, however, is of vital importance for the success or failure of penalty corners as the decision for an offensive play call that matches well with the opponent's defense strategy significantly increases the success rates of penalty corners (cf. Vinson et al., 2013).

Because of the explorative nature of this research, a questionnaire was developed in Study 1 which was designed to collect information about penalty corners from the perspective of expert players. That is, we mainly wanted to get basic information on the current state of penalty corners while also finding out how expert players think about application of a defense-dependent strategy in penalty corner situations, if they have already had some experience with these kinds of approaches and how they would describe their own gaze behavior and demands on their own gaze behavior (though reliability of self-reports on gaze behavior is limited, e.g., Kok et al., 2017; van Wermeskerken et al., 2018).

In the main study (Study 2), we scrutinized to what extent different strategical approaches are associated with different patterns of gaze and if strikers are able to distribute their allocation in a way that allows them to gain information on defenders' behavior on the one hand and focus on the execution of their own actions until the ball has reached the stopper, on the other. Furthermore, a fundamental difference between successful and less successful athletes across a range of different types of sports and sports situations is the ability to apply their visual perceptual skills in a targeted manner in order to be able to operate, anticipate, and react successfully (cf. Williams et al., 1999; Starks and Ericsson, 2003). Thus, Study 2 was

also aimed at investigating decision-making behavior (decision quality: optimal vs. less appropriate) as a function of gaze behavior during employing defense-dependent and -independent strategies in the penalty corner in field hockey.

STUDY 1

Study 1 was conducted as a preliminary study for Study 2. Given the explorative nature of this current research, it was important to generate subjective assessments about experts' behavior during penalty corners. In general, the following questions were answered by the participants through the questionnaire: (1) What is the importance of penalty corners during training for professional teams? (2) Which offense and defense tactics are preferable in professional teams and do they validate our initial assumptions? (3) How do experts behave during penalty corners from a tactical perspective? (4) How can defense-dependent strategies be implemented, i.e., where and when do players think one should gain information on the defenders' strategic behavior?

Method

Participants

In total, 48 (31 male, 17 female) participants completed the questionnaire. About 19% of all participants actively played field hockey in the highest German league and 81% in the second highest league at the time of data collection. About 17% of the participants reported having <1 year of playing experience in the two highest German leagues, about 50% had 1–5 years, 25% 5–10 years, and 8% more than 10 years of playing experience.

Materials and Procedure

The questionnaire[2] included 29 questions and was created online using the EFS Survey program (Questback GmbH, Germany). First, questions were asked about the preparation of penalty corners for the specific match in order to find out about the importance of penalty corner training in general. The survey also indicated whether the implementation of defense-dependent strategies seems promising from a player's perspective. Second, questions were asked about the theoretical preparation. Subsequently, participants answered questions concerning their (gaze) behavior *during the match*. Here the focus was on assessing individual gaze strategies and possible gaze strategies in defense-dependent approaches.

Results

Forty-eight participants indicated that they performed a penalty corner training at least once a week at the time of the data collection. More than half of the participants reported to train for penalty corners in a second session per week in addition to their regular practice. More than 70% of the participants took part in at least one training session specifically designed to improve penalty corner performance. Almost 80% of the respondents indicated that a penalty corner training during a team training unit usually lasted 15–30 min. The duration of the additional penalty corner training was around 30–45 min for approximately half of the respondents.

Players indicated that the penalty corner drag-flick is indeed the variation they most often choose and train for, followed by the 90° and deflection variations. Almost all the participants prepared for these variations using video analyses of the players from opposing teams. They indicated that countering the defense variation 3:1 3:1 with the 90° variation is the most appropriate solution, while the efforts do not match well with the 3:1 2:2 variation. They also supported our initial assumption that the deflection variation is optimal for countering the 3:1 2:2 variation, while being largely ineffective for the 3:1 variation.

Usually, with prior training, players decided on the variation to use either before the match or just before the penalty corner based on the knowledge of the tendencies of the opposing team. Though it was found that approximately 40% of the participants, at least once tried to adapt their strategical approach to the run-up behavior of the defensive players during the penalty corner, players also indicated that in the vast majority of the cases, they employed a defense-independent strategy (**Figure 1**).

With regard to the individual gaze strategies of the striker, 60% of the participants indicated that they had consulted with their coach about gaze behavior before. Regarding the time when gaze could be directed at the defenders to identify their strategy, "just after the ball was injected," and "just before the ball is stopped" were the most common responses by the participants (**Figure 2**). They emphasized that in these situations, but not while drag-flicking, there usually would be sufficient time to react to the run-out of the defenders, i.e., the preparation of the variation and the movement execution as a response to the strategy of the defenders. Unsurprisingly, most of the players named parts of the defense ("position trailer," "position first runner") or the "whole defense block" as information rich areas they would focus on to identify defensive strategies. Moreover, they named the deflection and 90° variations as the most suitable variations to react to the defenders' behavior.

Discussion

The results of the questionnaire indicated that there are frequent and regular penalty corner trainings in the first German division in field hockey. With a total duration of 2 h of training and an average of 15–30 min of specific penalty corner training, almost a quarter of the training is used exclusively for penalty corners. Many teams perform an additional penalty corner training to improve their success rate. In order to be prepared for the opponent, penalty corners as well as the opponent's team tactics are discussed during competition preparation. Based on these discussions, the penalty corner training is adapted and the different variants are determined before the actual match or penalty corner. The results of the questionnaire also illustrate that the offense variation is chosen shortly before the penalty corner in most cases, i.e., seemingly, the defense-independent strategy is employed much more often than the defense-dependent strategy even though the participants' level of expertise was relatively high. This may help to explain why in the past years, many teams have been able to lower the success rates of very good strikers through planned and purposeful defense behavior: they were able

[2]The German questionnaire can be found in the **Supplementary Material**.

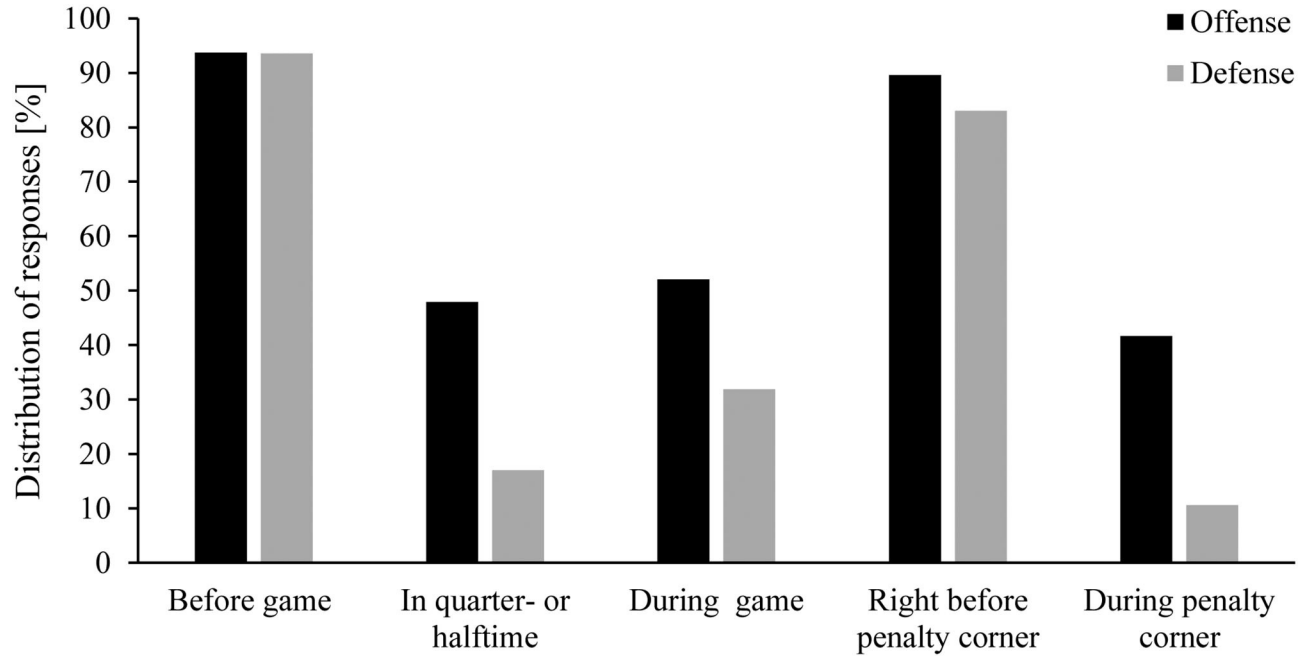


FIGURE 1 | Distribution of answers to the question when a decision is made for a certain penalty corner variation during competition by offensive players and defensive players. Participants were allowed to give multiple answers.

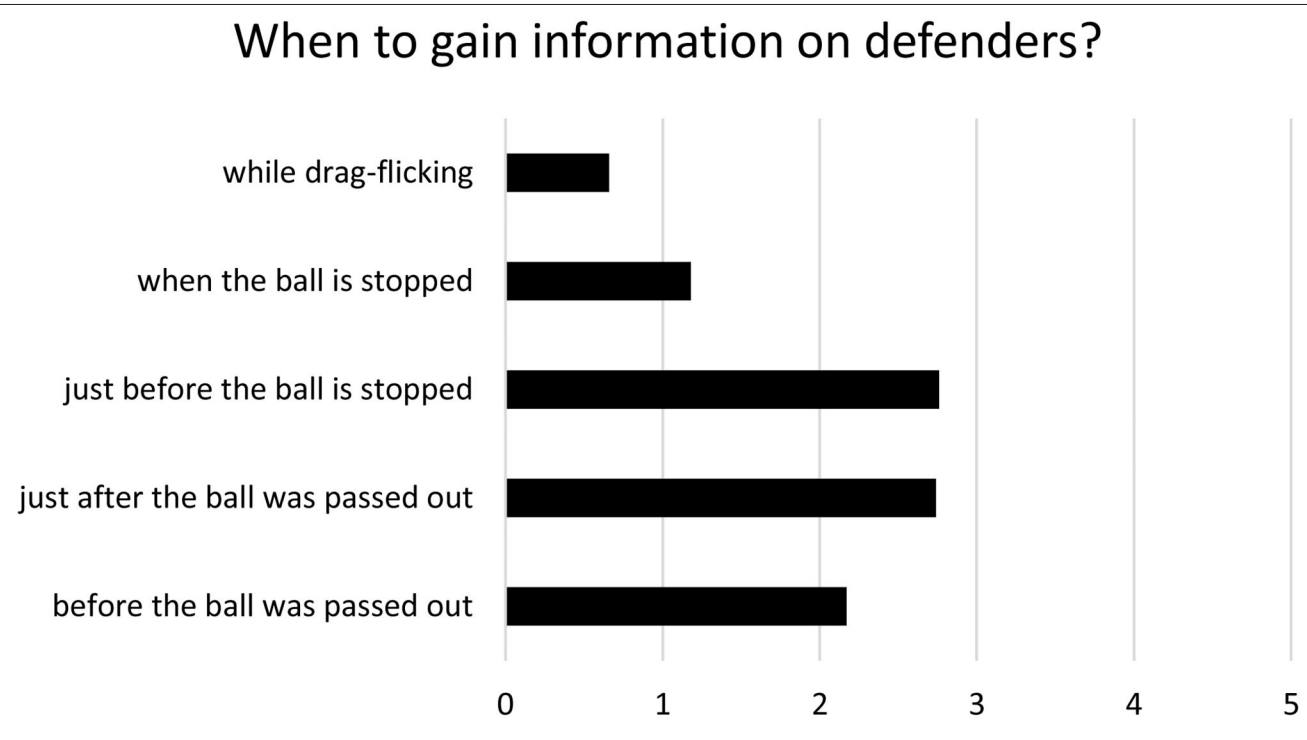


FIGURE 2 | Hockey players' opinion on when it is reasonable to acquire information on the defenders' run-out behavior/formation during penalty corners on a five-point-scale (1 = not possible/reasonable at all, 5 = very well possible/reasonable) in order to react to it by choosing a well-matching offensive variant.

to know what to expect and tried to employ a more promising counter strategy. If offense players more often employed defense-dependent decision making instead of the currently prevailing defense-independent decision making, it could help to select the offense strategy that fits best with the defenders' strategy. This would make it nearly impossible to prepare an effective counter

strategy for defense players as well. The players themselves considered this as a promising development. They indicated that the pass to the left (90° variation) and the deflection variation are the most suitable for short-term changes during actual penalty corner attempts, supporting our initial assumptions on which areas of interest may play a more important role in defense-dependent compared to defense-independent strategy.

STUDY 2

Based on the results of Study 1, and because self-reports on gaze behavior are not always reliable (cf. Kok et al., 2017; van Wermeskerken et al., 2018), Study 2 was designed as a field study with the goal of analyzing gaze and decision-making behaviors of strikers in field hockey. We aimed to scrutinize if (and how far) the gaze behavior differs between attempts in which the players know which variant they will play and attempts in which they try to counter the defending team's tactical approach (defense-dependent vs. defense-independent strategy). Thereby, the defense-independent strategy may be considered the players' natural (normally employed) strategy, based on what they reported in Study 1. In line with the information provided by the players in Study 1, we hypothesized that identifying tactical approaches of the defense requires looking longer at the goal area, especially early on (i.e., roughly until the ball reaches the stopper). This is so because later, the attention has to be directed to the stopper and the ball to guarantee optimal execution while possibly also perceiving information from other areas of interest (e.g., the stopper) peripherally (cf. Hüttermann et al., 2013; Klostermann et al., 2020). In contrast, while employing a defense-independent strategy, strikers theoretically can exclusively focus on offense related aspects of the penalty corner, such as the injecting player, stopper, and ball, early on because regardless of the defenders' run-out they would follow the same action plan anyway. So, there is a much lesser need to gain information on the defenders' behavior at that point. Furthermore, we also wanted to test if successfully identifying the defenders' tactical approach benefits from certain patterns of gaze behavior. We analyzed if and in how far gaze behavior during attempts in which players tried to identify and react to the defender's behavior differs between successful and unsuccessful trials. We expected that relatively more time will be spent looking at the seemingly more informative areas of interests (i.e., goal area and ball) in the cases the strikers made an optimal decision compared to an appropriate decision. Because of the results of Study 1 and also because several areas of interest (the injector, stopper, and goal area) are spatially far apart, we did not expect strikers to focus on areas in between until the ball had reached the stopper.

Method

Participants

In total, 14 strikers (3 female, 11 male) took part in the experiment. Due to technical difficulties, the eye-tracking data of one participant could not be evaluated and had to be excluded. The average age of the participants was 21.93 years ($SD = 3.95$ years). At the time of the experiment, the participants had been active as field hockey players for 16.71 years on average ($SD = 2.53$ years). Six of the participants (3 female, 3 male) had experience in the first German Division ($M = 3.33$; $SD = 2.34$). Nine of the players indicated experience in the second German Division ($M = 2.67$; $SD = 2.29$). Two of the participants also had experience as a striker in senior national teams (5 years) and another three where part of a youth national team for 1.67 years on average ($SD = 0.58$). The experiment was carried out in accordance with the Helsinki Declaration of 1975, and the participants signed a consent form approved by the local ethics committee.

Materials

The gaze behavior of each participant was recorded using a mobile eye-tracking system (Pupil Labs GmbH, Berlin, Germany). A mobile eye-tracking headset connected to a mobile bundle consisting of a Motorola Moto Z2 or Z3 Play with an USBC-USBC cable was used. The two eye cameras of the eye tracker had to be configured to record the full scope of the movement of the pupil in all movement directions. The front camera of the eye-tracking system had to be adjusted so that the entire visual field of the striker was recorded (120 frames per second). The gaze information of both eyes was recorded at 200 Hz and matched with a simultaneously captured scene video recorded at 30 Hz.

The game situations were recorded by two cameras (GoPro Hero 8 black, GoPro, Sam Mateo) from behind the striker and from a lateral perspective, to be able to view and assess all movements from two different viewing positions.

Procedure

The testing of each participant took around 30 min including the instruction, a warmup, and the actual testing. Initially, the test setup and procedure were explained, and the eye-tracking glasses were adjusted for each participant individually. After configuring the mobile recording devices (Moto Z2 or Z3 Play with Pupil Mobile App), they were connected, and the calibration was performed prior to the start of the testing.

Each participant performed 20 penalty corners as the striker (see **Figure 3**). In half of these penalty corners, the penalty corner variations were given beforehand (defense-independent strategy). The combination of penalty corner variations (shot, 90°, or deflection) were chosen in random order. In the other half players were asked to react to the run-out behavior of defenders (defense-dependent strategy). Importantly, defenders' strategy was always unknown to the offense and thus had to be identified during the penalty corner attempt (see **Supplementary Table 1**). Each defense variation was played five times in random order (5x 3:1 3:1, 5x 3:1 2:2) and defense variations were kept the same between both conditions. The striker's task was to find an optimal solution for the situation. To this end, three solution possibilities were available: penalty corner drag-flick, 90° variation, and deflection variation but it was emphasized that only the latter two were considered optimal solution depending on the run-out behavior of the defenders whereas the direct shot (penalty corner drag-flick) was considered a fallback option (which is never completely inappropriate to use).

FIGURE 3 | Depiction of field testing from the perspective behind the striker showing the striker (standing) and stopper of the offensive team in the foreground and the defenders and goalkeeper inside the goal before running out just after the injector (far left from the goal) has played the ball.

The experimental differentiation between the defense-dependent and defense-independent situations served to distinguish varying demands on gaze behavior. That way, both visual sources of information for the successful execution of an action (shot, pass) as well as sources of information for making correct decisions in the penalty corner situation were meant to be identified.

Data Analysis

First, the recordings of those cameras that recorded the individual shots were sifted through entirely. The video recordings served to identify the running behavior of the defense and, thereby, reconstruct the participants' decision-making behavior. The execution of the penalty corners was divided based on the decision quality (optimal vs. less appropriate) in this context by choosing the 90° variation as the optimal solution for 3:1 3:1 and the deflection variation as the optimal solution for 3:1 2:2. Respectively, the deflection variation was less appropriate for the 3:1 3:1 defensive variation and the 90° variation was less appropriate for the 3:1 2:2 variation.

Next, the video material of the eye-tracking system was calibrated offline in order to monitor the visual foci of the participants during the execution of the penalty corners. A manual frame-by-frame analysis was used to analyze the strikers' gaze behavior using the software Kinovea (Version 0.8.15; for a similar procedure, see Fasold et al., 2018). We only focused on the analysis of gaze duration and left out analyses of other gaze parameters (but see Di Nocera et al., 2007; Noël and van Der Kamp, 2012). Thereby, as common in velocity algorithms (Holmqvist et al., 2011) we included smooth pursuits of moving areas of interest, for example the ball, in our count of gaze durations (cf. Dicks et al., 2010; Aksum et al., 2020). Only if an area of interest was focused for 4 consecutive frames (120 ms) this was counted as gaze at a certain location. A second rater rated 10% of the trials in order to gain information on the reliability of the first rater's work. Cohen's Kappa was found to be 0.77 ("substantial agreement," cf. Landis and Koch, 1977). In order to be able to draw conclusions about the gaze and decision-making behavior of the participants, areas of interest were defined for the penalty corners, which could be observed by the striker during the shooting process. The ball, the injector, the stopper, the goal area, the shooting players in 90° variation, and the deflection variation were defined as such areas of interest. The starting point of a scene was defined as the moment when the ball is injected, and the end of the shooting process was defined as the moment when the ball has been passed or shot directly by the striker. For a more detailed analysis of the scenes and also to allow testing whether differences between condition occur mainly during the initial phase of a penalty corner or not, the sequence of a penalty corner was divided into three phases: Phase 1 is the period starting from the moment the ball is injected (at this point the defense is allowed to move) until the stopper's stick contact. Phase 2 has been marked as the period from the moment the ball leaves the stopper's stick until the striker receives the ball. Phase 3 ended at the moment the pass to the teammate was played or ball was shot (see **Figure 4**). Finally, data was analyzed using a 2 (condition: defense-dependent, defense-independent) × 3 (phase: phase 1, phase 2, phase 3) MANOVA with repeated measures for both factors and gaze duration at the different

FIGURE 4 | Pictures of a penalty corner attempt from phase 1 (top) to phase 3 (bottom). Phase 1 starts the moment the ball is injected and ends when the ball reaches the stopper's stick. Phase 2 has been defined as the period from the moment the ball leaves the stopper's stick until the striker receives the ball. Phase 3 ended at the moment the pass to the teammate was played or ball was shot.

areas of interest (ball, injector, stopper, goal area, deflecting player, 90° player) as dependent variables. Subsequently, data collected for the "defense-dependent" condition was analyzed by means of a 2 (decision quality: optimal; less appropriate) × 3 (phase: phase 1; phase 2; phase 3) × 2 (variation: deflection; 90°) MANOVA[3] with repeated measures for all factors and again gaze duration at the different areas of interest (ball, injector, stopper,

[3]Mean values that were submitted to both MANOVAs can be found in the **Supplementary Material**.

TABLE 1 | Summary of univariate analyses of gaze durations in defense-dependent and independent strategy for the three phases of the penalty corner process.

Effect	Dependent variable	F (dfs)	*p*	η_p^2
Condition	**Ball**	**14.035 (1, 12)**	**0.003**	**0.539**
	Injector	**17.923 (1, 12)**	**<0.001**	**0.599**
	Stopper	0.131 (1, 12)	0.724	0.011
	Goal area	**43.772 (1, 12)**	**<0.001**	**0.785**
	Deflecting player	0.99 (1, 12)	0.337	0.077
Phase	**Ball**	**9.898 (2, 24)**	**0.001**	**0.452**
	Injector	**41.773 (1.039, 12.473)**	**<0.001**	**0.777**
	Stopper	**36.539 (2, 24)**	**<0.001**	**0.753**
	Goal area	**59.372 (1.125, 13.497)**	**<0.001**	**0.832**
	Deflecting player	0.99 (2, 24)	0.383	0.077
Condition*Phase	**Ball**	**13.660 (2, 24)**	**<0.001**	**0.532**
	Injector	**10.736 (1.312, 15.740)**	**0.003**	**0.472**
	Stopper	2.022 (1.271, 15,256)	0.154	0.144
	Goal area	**52.095 (1.066, 12.786)**	**<0.001**	**0.813**
	Deflecting player	0.99 (2, 24)	0.337	0.077

These were used following a MANOVA in order to detect for which dependent variables differences exist. The columns including p values < 0.05 are in bold.

goal area, deflecting player, 90° player) as dependent variables. Assumptions for calculating a MANOVA were tested and in case of any violations of sphericity, Greenhouse-Geisser correction was applied. We followed both MANOVAs up with (univariate) ANOVAs in order to relate significant multivariate effects to single dependent variables.

Results

The data of a total of 253 penalty corners could be used for the analyses of gaze. Seven penalty corners could not be included in the data analysis due to technical problems. In total, the recordings consisted of 15,443 frames, of which 11,434 (74.04%) were included in the data analysis, as in 3,999 frames (25.9%), gaze was not detected. From Phase 1, there were 7,497 frames, of which 6,937 frames (92.53%) were included in the analysis. In phase 2, 4,047 frames were recorded, of which 2,426 frames (59.95%) showed no gaze marker. From the last phase, 2,080 frames (53.31%) of 3,902 frames could be considered for data analysis.

Differences Between Conditions

Results of the MANOVA showed a main effect of phase, $V = 1.761$; $F_{(10, 42)} = 30.981$, $p < 0.001$, $\eta_p^2 = 0.881$, of condition, $V = 0.873$; $F_{(5, 8)} = 10.964$, $p = 0.002$, $\eta_p^2 = 0.873$, and an interaction of phase and condition, $V = 0.955$; $F_{(10, 42)} = 3.835$, $p = 0.001$, $\eta_p^2 = 0.477$, on the participants' distribution of gaze on the different areas of interest.

Results of subsequent univariate analyses are shown in **Table 1**. Participants looked longer at the ball and injector ($M = 51.971$, $SE = 4.494$, 95% *CI* [42.179; 61.762]; $M = 5.832$, $SE = 0.743$, 95% *CI* [4.212; 7.452]) in the cases where a defense-independent strategy was adopted than in cases of a defense-dependent strategy ($M = 40.294$, $SE = 3.730$, 95% *CI* [32.166; 48.422]; $M = 3.493$, $SE = 0.607$, 95% *CI* [2.171; 4.816]). The opposite was true with regard to gaze durations on the goal area (defense-dependent: $M = 13.940$, $SE = 1.444$, 95% *CI* [10.793; 17.087]; defense-independent: $M = 3.722$, $SE = 0.859$, 95% *CI* [1.850; 5.595]).

However, while gaze duration at the ball during the defense-dependent and defense-independent conditions was roughly the same in phase 2 ($M = 39.267$, $SE = 6.020$, 95% *CI* [26.150; 52.384] vs. $M = 36.977$, $SE = 6.932$, 95% *CI* [21.873; 52.082]) and phase 3 ($M = 69.518$, $SE = 8.696$, 95% *CI* [50.571; 88.465] vs. $M = 68.914$, $SE = 9.023$, 95% *CI* [49.254; 88.575]), during phase 1, players focused on the ball longer in the defense-independent compared to the defense-dependent condition ($M = 47.127$, $SE = 5.785$, 95% *CI* [34.523; 59.732] vs. $M = 14.990$, $SE = 4.137$, 95% *CI* [5.976; 24.004]). In phase 1, players looked longer at the injector in cases they already knew how to carry out the penalty corner compared to the defense-dependent condition ($M = 16.237$, $SE = 2.258$, 95% *CI* [11.316; 21.157] vs. $M = 10.385$, $SE = 1.854$, 95% *CI* [6.346; 14.424]). However, this difference between conditions got smaller during phase 2 ($M = 1.260$, $SE = 0.702$, 95% *CI* [−0.271; 2.790] vs. $M = 0.095$, $SE= 0.095$, 95% *CI* [−0.112; 0.302]) and during phase 3, players in both the situations never looked at the injector. The gaze durations at the goal area during phase 1 ($M = 39.123$, $SE = 4.221$, 95% *CI* [29.925; 48.320] vs. $M = 9.276$, $SE= 2.478$, 95% *CI* [3,878; 14.674]) and phase 2 ($M = 2.697$, $SE = 1.044$, 95% *CI* [0.424; 4.971] vs. $M = 0.870$, $SE = 0.490$, 95% *CI* [−0 to 197; 1.937]) were longer in the defense-dependent compared to the defense-independent condition. But in phase 3, the goal area was only sparsely looked at in the defense-independent condition ($M = 1.021$, $SE = 1.044$, 95% *CI* [-0.846; 2.888]) (**Figure 5**).

The scoring rates of both conditions were rather similar. In the defense-dependent condition, the offense scored in 15.87% of the attempts whereas scoring rate was 16.54% in the defense-independent condition.

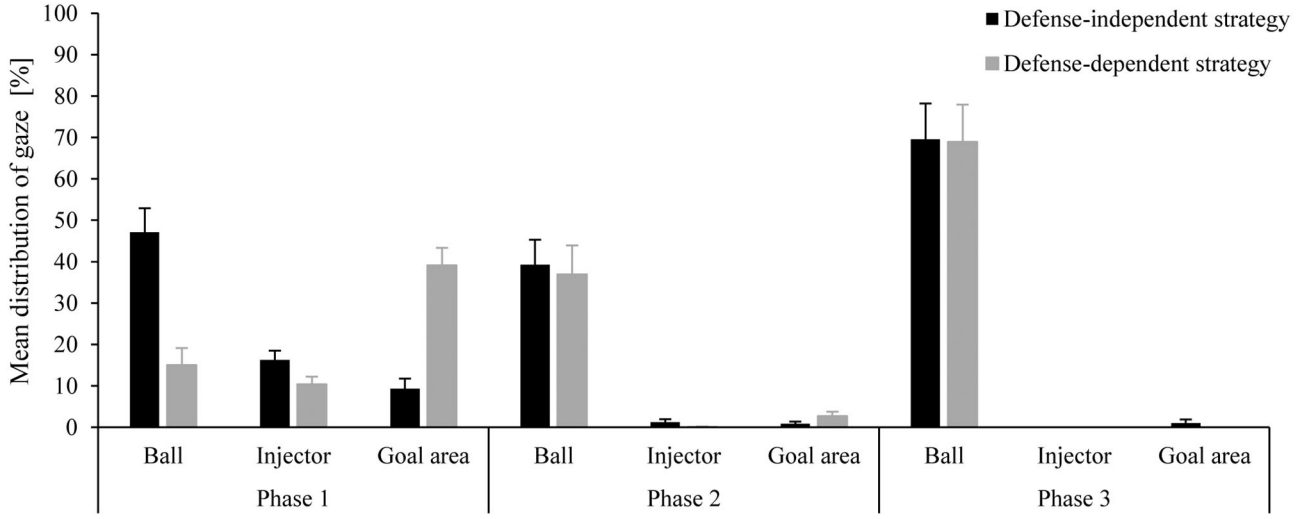


FIGURE 5 | Mean distribution (and SD) of gaze durations at the areas of interest (ball, injector, goal area) for both conditions (defense-independent; defense-dependent) as a function of time (Phase 1, Phase 2; Phase 3) and in relation to the length of a phase.

TABLE 2 | Summary of univariate analyses of gaze in defense-dependent strategy as a function of decision quality (optimal vs. less appropriate) and phase of the penalty corner.

Effect	Dependent variable	F (dfs)	*p*	η_p^2
Decision quality	**Ball**	**24.075 (1, 12)**	**<0.001**	**0.667**
	Injector	3.898 (1, 12)	0.072	0.245
	Stopper	**20.055 (1, 12)**	**0.001**	**0.626**
	Goal area	**25.116 (1, 12)**	**<0.001**	**0.677**
Decision quality*Phase	**Ball**	**9.319 (2, 24)**	**0.001**	**0.437**
	Injector	3.406 (1.007, 12.088)	0.089	0.221
	Stopper	**11.828 (2, 24)**	**<0.001**	**0.496**
	Goal area	20.216 (1.247, 14.960)	$p < 0.001$	0.628
Variant*Phase	Ball	0.191 (2, 24)	0.828	0.016
	Injector	**7.406 (1.034, 12.404)**	**0.017**	**0.382**
	Stopper	3.193 (1.034, 12.432)	0.09	0.21
	Goal area	0.795 (1.238, 14.858)	0.413	0.062

These were used following a MANOVA in order to detect for which dependent variables differences exist. The columns including p values < 0.05 are in bold.

Differences Within Defense-Dependent Condition

In 65.08% of the trials, the striker chose the optimal strategy to counter the defenders' run out. Results of the second MANOVA showed again that gaze differed between the phases of the penalty corner, $V = 1.556$; $F_{(8, 44)} = 19.296$, $p < 0.001$, $\eta_p^2 = 0.778$, but also that gaze was different for penalty corners in which players were able to respond to the behavior of the defense in an optimal compared to a less optimal way, $V = 0.756$; $F_{(4, 9)} = 6.971$, $p = 0.008$, $\eta_p^2 = 0.756$. Furthermore, there was an interaction between both factors, $V = 1.196$; $F_{(8, 44)} = 8.174$, $p < 0.001$, $\eta_p^2 = 0.598$, whereas gaze differences between phases weren't the same for deflection and 90° variations, $V = 0.595$; $F_{(8, 44)} = 2.327$, $p = 0.035$, $\eta_p^2 = 0.297$. Gaze did not differ between both penalty corner variants, though, $V = 0.421$; $F_{(4, 9)} = 1.634$, $p = 0.248$.

Results of subsequent, univariate analyses are shown in **Table 2**. In the cases where the players' behavior optimally matched the tactical formation of the defense, they spent more time looking at the ball ($M = 40.429$, $SE = 3.720$, 95% *CI* [32.324; 48.534] vs. $M = 16.187$, $SE = 5.183$, 95% *CI* [4.894; 27.481]; $p < 0.001$), the stopper ($M = 3.463$, $SE = 0.763$, 95% *CI* [1.800; 5.127] vs. $M = 7.597$, $SE = 2.118$, 95% *CI* [2.982; 12.212]; $p = 0.001$), and the goal area ($M = 14.399$, $SE = 1.437$, 95% *CI* [11.269; 17.530] vs. $M = 6.071$, $SE = 1.777$, 95% *CI* [2.199; 9.942]; $p < 0.001$).

However, these differences based on the appropriateness of the chosen tactical approach, appeared to be inconsistent across the different phases of the penalty corner (**Figure 6**). In the cases where the variant of the penalty corner matched the defensive formation optimally, players spent more time looking at the ball (optimal: $M = 13.376$, $SE = 4.411$, 95% *CI* [3.765; 22.987] vs.

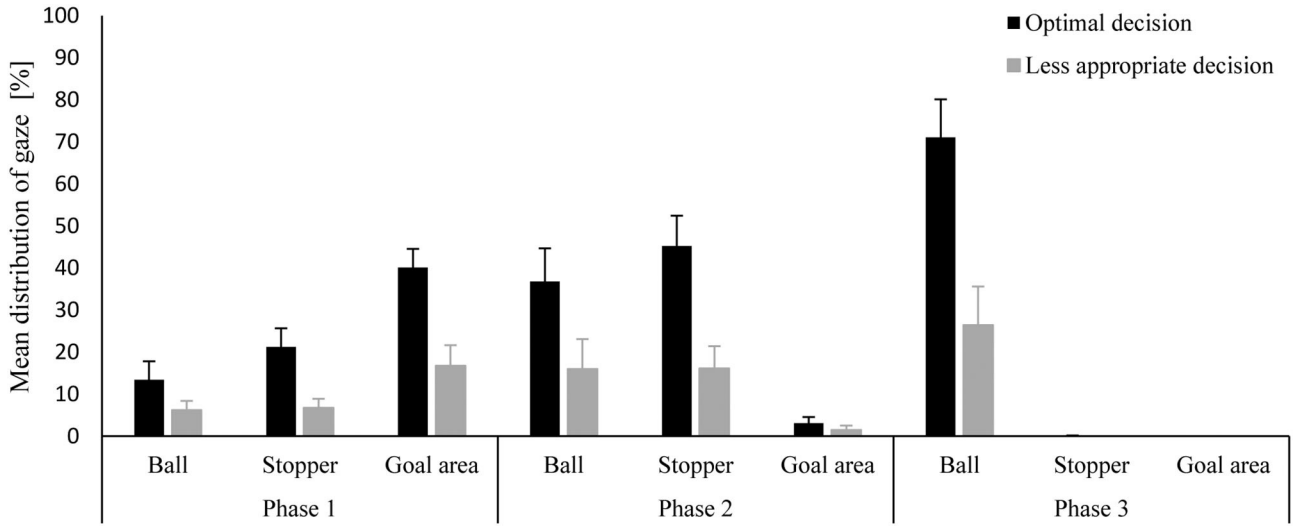


FIGURE 6 | Mean distribution (and SD) of gaze duration at the areas of interest (ball, stopper, goal area) for the evaluation of how well an offensive strategy has matched the run-out of defenders (optimal; less appropriate) as a function of time (Phase 1, Phase 2; Phase 3) and in relation to the length of a phase.

less appropriate: $M = 6.223$, $SE = 2.173$, 95% CI [1.490;10.957]), the stopper (optimal: $M = 21.245$, $SE = 4.429$, 95% CI [11.596; 30.894] vs. less appropriate: $M = 6.702$, $SE = 2.179$, 95% CI [1.955; 11.448]), and the goal area (optimal: $M = 40.127$, $SE = 4.407$, 95% CI [30.525; 49.728] vs. less appropriate: $M = 16.717$, $SE = 4.924$, 95% CI [5.988; 27.445]) in phase 1 than in the cases where the chosen play design was less appropriate.

Regarding gaze spent at the ball, this difference got bigger in phase 2 (optimal: $M = 36.817$, $SE = 7.848$, 95% CI [19.718; 53.915] vs. less appropriate: $M = 15.962$, $SE = 7.116$, 95% CI [0.457; 31.466]), and was biggest in phase 3 (optimal: $M = 71.095$, $SE = 9.040$, 95% CI [51.398; 90.793] vs. less appropriate: $M = 26.377$, $SE = 9.221$, 95% CI [6.286; 46.468]). This pattern was different for other areas of interest. Although, the differences in gaze duration for the stopper were biggest in phase 2 (optimal: $M = 45.260$, $SE = 7.177$, 95% CI [29.622; 60.897] vs. less appropriate: $M = 16.090$, $SE = 5.295$, 95% CI [4.552; 27.627]), players hardly looked at the stopper in the third phase. Differences in relation to gaze durations for the goal area disappeared after phase 1 (after which the goal area was hardly focused on at all).

Discussion

In Study 2, the gaze behavior of strikers was recorded in defense-dependent and -independent conditions. The results show that both strategies differ in important aspects of their gaze behavior, mainly in phase 1. This is in line with the descriptions of gaze behavior of players in Study 1 which was unexpected because the self-reports on gaze are not always reliable. In this case, however, players' temporal and spatial description of acquiring information on defenders' behavior matched well with what we actually observed in Study 2. This pattern of gaze allowed players to choose the strategy that matched the run-out of defenders optimally in almost two out of three penalty corners. Players in the defense-dependent condition appear to spend more time looking at the goal area at the expense of gaze durations on the ball and injector (until the ball has reached the stopper and the strikers have to initiate their own actions). In fact, the ball was fixated considerably less (∼15%) compared to the defense-independent condition (∼47%). In the defense-dependent condition, on the other hand, the goal area was fixated longer (∼39%) than in the defense-independent condition (∼9%). That is, as hypothesized players in the defense-dependent condition had to manage two sources of information, one providing information on the defense (i.e., which variation to choose) and one providing information that seemed necessary for more optimal execution of their following actions (shot, pass). In contrast, players in the defense-independent strategy knew the offensive variation already before the injection of the ball what allowed them to mainly focus on the ball (and stopper) throughout the initial course of the penalty corner attempt. Rothkopf et al. (2007) emphasized that areas of high interest are focused on at the beginning of a phase or a sequence. But probably because the players in the defense-dependent condition needed sufficient time to not only perceive the pattern of defenders' run-out behavior, but also to react to it in an appropriate way, differences between conditions mainly vanished after phase 1. However, during the penalty corner process, strategies did not lead to different patterns of gaze because the focus shifted predominantly on movement execution.

In phase 2, this involves gaze at the stopper who indicates the striker where to pick up the ball, whereas in phase 3, gaze was almost exclusively directed at the ball to guide the execution of strikers' actions. The latter is also supported by a study by Kurz et al. (2018), which showed that focusing the ball is important for a good technical execution. Similar to the current interaction finding Noël and van Der Kamp (2012) revealed that penalty takers paid more visual attention

to the goalkeeper in the beginning of their run-up when they were asked to employ a keeper-dependent strategy. Importantly, they also stopped collecting information on the goalkeeper's behavior (jump direction, movement onset of dive) before foot-to-ball contact during their run-up because it would not have been possible to consider very late sources of information on the goalkeeper and successfully react to them anyway. Taken together, the current results indicate that participants indeed tried to react to the defense using adapted gaze strategies to decide on early parts of the defenders run-out behavior (cf. Roth et al., 2007). However, the current results also show that it probably will not be possible to take later aspects of defenders' run-out into account.

Furthermore, the differentiation between gaze behaviors when decision making was considered optimal or less appropriate during the defense-dependent condition, points to the importance of three areas of interest for adaptive decision making: the goal area, the stopper, and the ball. When choosing the optimal variant to counter defenders' behavior, the strikers spent more time looking at each of these areas. Probably, the longer gaze times on the ball and stopper, mainly in the second and third phase respectively, are a consequence of having identified the defenders' strategy in time (in phase 1) after which they can solely focus on the execution of their own actions. In case they have more problems recognizing defenders' strategies, they probably still have to focus on other potentially informative areas before shifting their focus to guidance of their own movements. Spending sufficient time on the goal area, though, seems to be most the important factor during phase 1 to recognize the defenders' behavior correctly. If the strikers chose the optimal strategy, they would spend more than twice as much time looking at the goal area than in the cases where they chose the less appropriate option.

In this current study, participants almost never spent time looking in between areas of interest. This is probably related to the fact that some of the areas of interest are very far apart and also a consequence of the decision to not differentiate between sources of information in the goal area. It seems likely that within this area of interest, participants made use of peripheral vision to perceive several defenders, the goalkeeper, and the target area in the goal, at the same time (cf. Hüttermann and Memmert, 2017). However, given the relatively long distance between the striker and the goal, the short distance between defenders and the fact that the goal and the goalkeeper were right behind them, it seemed impossible to reliably differentiate between gazes to these sources of information. It remains a question for future research, though, to examine the extent to which strikers make use of peripheral vision, especially during defense-dependent strategy where the need to perceive different sources of information seems stronger (cf. Hüttermann et al., 2014).

Furthermore, the strikers were not trained to employ defense-dependent strategy though. That is why it seems reasonable that gaze behavior in this condition would potentially look somewhat different after players/teams have gathered more experience with this strategy. However, despite this fact, the strikers were able to choose an optimal strategy in 65.08% of the cases. This strongly points to adaptive offensive play calling (i.e., providing offenses with at least two variants of which they can choose based on how the defenders run out) as a promising future development to improve penalty corner success. This is supported by the players' self-evaluation provided in Study 1. Furthermore, goal scoring rate was similar in both conditions, but comparisons of percentages of goals scores seems rather problematic. First, the focus of the current study was mainly on gaze behavior and decision making of the striker. However, goal scoring does not only depend on his/her decision making but also on the perception and performance of his/her teammates (especially in case he/she opts to pass and not to shoot directly). Second, goal scoring does not only depend the defense tactical approach but also on other aspects of the players' behavior and performance (cf. Vinson et al., 2013). For instance, a similar shot on the goal will sometimes result in a goal, but sometimes be saved by the goalkeeper. Finally, as stated above, the strikers and also the other offensive players were not trained to adapt their behavior during a penalty corner. That is, goal scoring rate would probably increase after proper training sessions.

GENERAL DISCUSSION

In recent years, various studies have dealt with different gaze strategies whose application is meant to benefit players' decision making and performance in sports games (e.g., Wilson et al., 2015). However, patterns of gaze behavior likely differ between different sports and probably also within different situations/tasks within one sport (Cañal-Bruland and Mann, 2015). This is so because transfer of knowledge in one sport (e.g., gaze behavior in soccer penalty kicks, e.g., Wilson et al., 2009) is not easy and therefore, cannot replace research in other sports, such as field hockey, that have received less attention by sport scientists/sport psychologists in the past. However, the general principles and observations from one sport can indeed help to improve or better understand decision-making behavior in another sport. In the current paper, we tried to scrutinize if and to what extent can strikers in hockey penalty corners also consider the actions of their opponents (as e.g., in soccer penalty kicks, cf. van der Kamp, 2006, 2011). We asked how far a reaction (to an opponent) is be more effective than a self-initiated, already planned action. To this end, previous research on the relationship between action and reaction has mainly focused on movement times (e.g., Welchman et al., 2010). Those findings indicate that reactive movements are usually faster than self-initiated movements (Pinto et al., 2011) and that this holds across different levels of within-task expertise (Martinez de Quel and Bennett, 2014). However, there also seem to be other benefits of choosing to react to an opponents' behavior (cf. Noël et al., 2021). It allows to ultimately choose a strategy that is more promising because decisions are based on more (reliable) information of the opponents' behavior. In contrast, it causes extra problems like time constraints, the need to synchronize more complex processes as a team, and additional demands on gaze behavior, too.

The present study was focused on the investigation of the gaze behavior and decision making of experienced field hockey

players during penalty corners. We were interested in how far the offensive players can adapt their strategy to the run-out behavior of the defenders, thus in how far it is reasonable to base their own actions on the perceptions of the opponents' behavior. In Study 1, a questionnaire was used with the goal of obtaining subjective experiences with and opinions on adaptive offensive behavior during penalty corners and its associated gaze behavior mainly in order to subsequently examine gaze behavior of defense-dependent and -independent strategy within the scope of a field study (Study 2). That was necessary because of the explorative nature of the current research and missing information on many basic relationships in this context.

Both the studies together show that adapting to the behavior of the defenders seems possible and is considered a promising future development by most players. Furthermore, a look at the decision-making performance of one of the players with experience as striker in the German senior national team illustrates that especially high-class players after more intensive training are very well capable of reacting to the defenders' run-up in the first phase of the penalty corner. This particular participant always chose the optimal tactic to counter the approach of the defense.

However, it remains to be seen in what way adaptive behavior during penalty corners can be trained because rapid reactions to the defenders run-out do not only afford good decision making by an individual but also communication between offensive players and coordination of their gaze behavior and movements (cf. Fasold et al., 2018, 2021). That is, after strikers have learnt/established a certain pattern of gaze behavior over the course of several training sessions, they are probably able to focus on the right place at the right time (cf. Magill, 1998) and know the more informative areas enabling good decision making (Abernethy, 2001). But implementing of clear arrangements concerning the routes and positions of the other offensive players and how they get informed on the strikers' final decision seems to be al longer learning process. Furthermore, it certainly requires very good technical skills of all attackers to allow for error-free employment of defense-dependent strategy.

Taken together, the current results point to the benefits of employing a defense-dependent strategy (or in more general terms: reacting instead of acting completely self-planed) at least from time to time also to keep defenses uncertain about which variations they should expect. However, employment of such a strategy seemingly requires intense training and a certain skill set among the offensive players allowing them to rapidly change and coordinate their behavior. Furthermore, on a more strategical level, play designers have to determine out of which more specific variations strikers should chose while observing the run-out of defenders. That is, the current study can be considered a first step toward implementing adaptive decision making by the offense, but there is much work left for coaches, players, and researchers to find out under which circumstances defense-dependent strategies work best. For example, it remains to be investigated for future scientific work to what extent the analysis of other parameters of gaze behavior can be used. In this context, it would also be interesting to analyze to what extent the current results can be replicated (if for example it is made use of a dispersion-based algorithm to identify fixations, see e.g., Blignaut and Beelders, 2009) or to what extent other gaze data can support the current results.

Nevertheless, adaptation to the defenders' formation and behavior during game play is also found in other sports as American Football when, for example, a receiver modifies his/her route according to previous instructions based on his interpretation of the defense strategy. Though learning how to adapt during penalty corners appears relatively extensive, employing defense-dependent strategies seems very well implementable. This appears to also be the case in other sports in which a player or a team has to decide between self-initiated actions and waiting for an action of the opponent in order to choose a reaction that matches the opponents' behavior well.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by German Sport University Cologne. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SK, BN, and FF developed the study concept and contributed to the design. AS and LH collected the data. SK and BN wrote the first draft of the manuscript. All authors helped to edit and revise the manuscript and approved the final submitted version of the manuscript.

FUNDING

This study was funded by the Federal Institute of Sports Science (Bundesinstitut für Sportwissenschaft).

ACKNOWLEDGMENTS

The authors acknowledge financial support by the German Research Foundation and the University of Rostock within the funding program Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.674511/full#supplementary-material

REFERENCES

Abernethy, B. (2001). "Attention," in *Handbook of Sport Psychology*, eds R. Singer, H. Hausenblas, and C. Janelle (New York, NY: Wiley and Sons), 53–85.

Aksum, K. M., Magnaguagno, L., Bjørndal, C. T., and Jordet, G. (2020). What do football players look at? An eye-tracking analysis of the visual fixations of players in 11 v 11 elite football match play. *Front. Psychol.* 11:562995. doi: 10.3389/fpsyg.2020.562995

Baker, J., Farrow, D., Elliott, B. C., and Anderson, J. (2009). The influence of processing time on expert anticipation. *Int. J. Sport Psychol.* 40, 476–488.

Blignaut, P., and Beelders, T. (2009). The effect of fixational eye movements on fixation identification with a dispersion-based fixation detection algorithm. *J. Eye Mov. Res.* 2, 1–14. doi: 10.16910/jemr.2.5.4

Cañal-Bruland, R., and Mann, D. L. (2015). Time to broaden the scope of research on anticipatory behavior: a case for the role of probabilistic information. *Front. Psychol.* 6:1518. doi: 10.3389/fpsyg.2015.01518

Di Nocera, F., Camilli, M., and Terenzi, M. (2007). A random glance at the flight deck: pilots' scanning strategies and the real-time assessment of mental workload. *J. Cogn. Eng. Decis. Mak.* 1, 271–285. doi: 10.1518/155534307X255627

Dicks, M., Button, C., and Davids, K. (2010). Examination of gaze behaviors under *in situ* and video simulation task constraints reveals differences in information pickup for perception and action. *Attent. Percept. Psychophys.* 72, 706–720. doi: 10.3758/APP.72.3.706

Fasold, F., Nicklas, A., Seifriz, F., Schul, K., Noël, B., Aschendorf, P., et al. (2021). Gaze coordination of groups in dynamic events - A tool to facilitate analyses of simultaneous gazes within a team. *Front. Psychol.* 12:656388. doi: 10.3389/fpsyg.2021.656388

Fasold, F., Noël, B., Wolf, F., and Hüttermann, S. (2018). Coordinated gaze behaviour of handball referees: a practical exploration with focus on the methodical implementation. *Mov. Sport Sci. Sci.* 102, 71–79. doi: 10.1051/sm/2018029

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*. Oxford: Oxford University Press.

Hüttermann, S., and Memmert, D. (2017). The attention window: a narrative review of limitations and opportunities influencing the focus of attention. *Res. Q. Exerc. Sport* 88, 169–183. doi: 10.1080/02701367.2017.1293228

Hüttermann, S., Memmert, D., and Liesner, F. (2014). Finding the happy medium: an analysis of gaze behavior strategies in a representative task design of soccer penalties. *J. Appl. Sport Psychol.* 26, 172–181. doi: 10.1080/10413200.2013.816892

Hüttermann, S., Memmert, D., Simons, D. J., and Bock, O. (2013). Fixation strategy influences the ability to focus attention on two spatially separate objects. *PLoS ONE* 8:e65673. doi: 10.1371/journal.pone.0065673

Hüttermann, S., Noël, B., and Memmert, D. (2018). Eye tracking in high-performance sports: evaluation of its application in expert athletes. *Int. J. Comp. Sci. Sport* 17, 182–203. doi: 10.2478/ijcss-2018-0011

Ibrahim, R., Faber, G. S., Kingma, I., and van Dieen, J. H. (2017). Kinematic analysis of the drag flick in field hockey. *Sports Biomech.* 16, 45–57. doi: 10.1080/14763141.2016.1182207

Klatt, S., and Smeeton, N. J. (2021a). Attentional and perceptual capabilities are affected by high physical load in a simulated soccer decision-making task. *Sport Exerc. Perform. Psychol.* 10, 205–216. doi: 10.1037/spy0000228

Klatt, S., and Smeeton, N. J. (2021b). Processing visual information in elite junior soccer players: effects of chronological age and training experience on visual perception, attention, and decision making. *Eur. J. Sport Sci.* doi: 10.1080/17461391.2021.1887366. [Epub ahead of print].

Klostermann, A., Vater, C., Kredel, R., and Hossner, E. J. (2020). Perception and action in sports. On the functionality of foveal and peripheral vision. *Front. Sports Active Liv.* 1:66. doi: 10.3389/fspor.2019.00066

Kok, E. M., Aizenman, A. M., Võ, M. L.-H., and Wolfe, J. M. (2017). Even if I showed you where you looked, remembering where you just looked is hard. *J. Vis.* 17:2. doi: 10.1167/17.12.2

Kredel, R., Vater, C., Klostermann, A., and Hossner, E. J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: a systematic review of 40 years of research. *Front. Psychol.* 8:1845. doi: 10.3389/fpsyg.2017.01845

Kuhn, W. (1988). "Penalty-kick strategies for shooters and goalkeepers," in *Science and Football,* eds T. Reilly, A. Lees, K. Davids and W. J. Murphy (London: E and F.N Spon), 489–492.

Kurz, J., Hegele, M., and Munzert, J. (2018). Gaze behavior in a natural environment with a task-relevant distractor: how the presence of a goalkeeper distracts the penalty taker. *Front. Psychol.* 9:19. doi: 10.3389/fpsyg.2018.00019

Laird, P., and Sutherland, P. (2003). Penalty corners in field hockey: a guide to success. *Int. J. Perform. Anal. Sport* 3, 19–26. doi: 10.1080/24748668.2003.11868270

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159. doi: 10.2307/2529310

Magill, R. A. (1998). Knowledge is more than we can talk about: implicit learning in motor skill acquisition. *Res. Q. Exerc. Sport*. 69, 104–110. doi: 10.1080/02701367.1998.10607676

Martinez de Quel, O., and Bennett, S. J. (2014). Kinematics of self-initiated and reactive karate punches. *Res. Q. Exerc. Sport*. 85, 117–123. doi: 10.1080/02701367.2013.872222

Morris-Binelli, K., Müller, S., van Rens, F. E., Harbaugh, A. G., and Rosalie, S. M. (2021). Individual differences in performance and learning of visual anticipation in expert field hockey goalkeepers. *Psychol. Sport Exerc.* 52:101829. doi: 10.1016/j.psychsport.2020.101829

Morris-Binelli, K., van Rens, F. E., Müller, S., and Rosalie, S. M. (2020). Psycho-perceptual-motor skills are deemed critical to save the penalty corner in international field hockey. *Psychol. Sport Exerc.* 51:101753. doi: 10.1016/j.psychsport.2020.101753

Noël, B., Furley, P., Hüttermann, S., Nopp, S., Vogelbein, M., and Memmert, D. (2014). Einflussfaktoren auf Erfolg und Misserfolg beim Elfmeterschießen: Eine empiriegeleitete retrospektive Analyse der Europa- und Weltmeisterschaften von 1982 bis 2012. *Z. Sportpsychol.* 21, 51–62. doi: 10.1026/1612-5010/a000118

Noël, B., and van Der Kamp, J. (2012). Gaze behaviour during the soccer penalty kick: an investigation of the effects of strategy and anxiety. *Int. J. Sport Psychol.* 43:326.

Noël, B., van der Kamp, J., and Klatt, S. (2021). The interplay of goalkeepers and penalty takers affects their chances of success. *Front. Psychol.* 12:645312. doi: 10.3389/fpsyg.2021.645312

Piñeiro, R., Sampedre, J., and Refoye, I. (2007). Differences between international men's and women's teams in the strategic action of the penalty corner in field hockey. *Int. J. Perform. Anal. Sport* 7, 67–83. doi: 10.1080/24748668.2007.11868411

Pinto, Y., Otten, M., Cohen, M. A., Wolfe, J. M., and Horowitz, T. S. (2011). The boundary conditions for Bohr's law: when is reacting faster than acting? *Attent. Percept. Psychophys.* 73, 613–620. doi: 10.3758/s13414-010-0057-7

Rosalie, S. M., McIntyre, A. S., Stockman, S., King, C., Watkins, C., Wild, C. Y., et al. (2017). Does skill specialisation influence individual differences in drag flicking speed and accuracy? *J. Sports Sci.* 35, 602–609. doi: 10.1080/02640414.2016.1180422

Roth, K., Schorer, J., and Peters, B. (2007). *Blickbewegungsdiagnostik bei Hockeynationalspielern und ihre Trainingsimplikationen. BISp-Jahrbuch: Forschungsförderung 2006/2007*. Bonn: Statistisches Bundesamt, 241–244.

Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2007). Task and context determine where you look. *J. Vis.* 7, 16.1–20. doi: 10.1167/7.14.16

Rowe, S. (2019). *Complete goalscoring analysis: Men's Hockey World Cup 2018*. Unpublished manuscript.

Starks, J., and Ericsson, K. (2003). *Expert Performance in Sports: Advances in Research on Sport Expertise*. Champaign, IL: Human Kinetics.

van der Kamp, J. (2006). A field simulation study of the effectiveness of penalty kick strategies in soccer: late alterations of kick direction increase errors and reduce accuracy. *J. Sports Sci.* 24, 467–477. doi: 10.1080/02640410500190841

van der Kamp, J. (2011). Exploring the merits of perceptual anticipation in the soccer penalty kick. *Motor Control* 15, 342–358. doi: 10.1123/mcj.15.3.342

van Wermeskerken, M., Litchfield, D., and van Gog, T. (2018). What am I looking at? Interpreting dynamic and static gaze displays. *Cogn. Sci*. 42, 220–252. doi: 10.1111/cogs.12484

Vater, C., Kredel, R., and Hossner, E.-J. (2016). Detecting single-target changes in multiple object tracking: the case of peripheral vision. *Attent. Percept. Psychophys*. 78, 1004–1019. doi: 10.3758/s13414-016-1078-7

Vater, C., Kredel, R., and Hossner, E.-J. (2017). Examining the functionality of peripheral vision: from fundamental understandings to applied sport science. *Curr. Issues Sport Sci*. 2:10. doi: 10.36950/2017ciss010

Vinson, D., Padley, S., Croad, A., Jeffreys, M., Brady, A., and James, D. (2013). Penalty corner routines in elite women's indoor field hockey: prediction of outcomes based on tactical decisions. *J. Sports Sci*. 31, 887–893. doi: 10.1080/02640414.2012.757341

Vizcaya, F. J. (2015). *Quantitative Analyse der Strafecken bei der Weltmeisterschaft 2014 in Den Haag. Ergebnisbericht*. Leipzig: IAT.

Welchman, A. E., Stanley, J., Schomers, M. R., Miall, R. C., and Bülthoff, H. H. (2010). The quick and the dead: when reaction beats intention. *Proc. R. Soc. B Biol. Sci*. 277, 1667–1674. doi: 10.1098/rspb.2009.2123

Williams, A., Davids, K., and Williams, J. (1999). *Visual Perception and Action in Sport*. London: E and F.N Spon.

Wilson, M. R., Causer, J., and Vickers, J. N. (2015). "Aiming for excellence: the quiet eye as a characteristic of expertise," in *Routledge International Handbooks. Routledge Handbook of Sport Expertise*, eds J. Baker and D. Farrow (London: Routledge/Taylor and Francis Group), 22–37.

Wilson, M. R., Wood, G., and Vine, S. J. (2009). Anxiety, attentional control, and performance impairment in penalty kicks. *J. Sport Exerc. Psychol*. 31, 761–775. doi: 10.1123/jsep.31.6.761

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 August 2021
doi: 10.3389/fpsyg.2021.681042

Gaze Behavior During Navigation and Visual Search of an Open-World Virtual Environment

Leah R. Enders[1], Robert J. Smith[1], Stephen M. Gordon[1†], Anthony J. Ries[2,3] and Jonathan Touryan[2†]*

[1] DCS Corp., Alexandria, VA, United States, [2] DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD, United States, [3] Warfighter Effectiveness Research Center, U.S. Air Force Academy, Colorado Springs, CO, United States

OPEN ACCESS

Edited by:
Frédéric Gosselin, Université de Montréal, Canada

Reviewed by:
Christoph Hoelscher, ETH Zürich, Switzerland
Otto Lappi, University of Helsinki, Finland
Vsevolod Peysakhovich, Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), France

***Correspondence:**
Leah R. Enders
lenders@dcscorp.com

† These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *15 March 2021*
Accepted: *28 June 2021*
Published: *09 August 2021*

Citation:
Enders LR, Smith RJ, Gordon SM, Ries AJ and Touryan J (2021) Gaze Behavior During Navigation and Visual Search of an Open-World Virtual Environment. Front. Psychol. 12:681042. doi: 10.3389/fpsyg.2021.681042

Eye tracking has been an essential tool within the vision science community for many years. However, the majority of studies involving eye-tracking technology employ a relatively passive approach through the use of static imagery, prescribed motion, or video stimuli. This is in contrast to our everyday interaction with the natural world where we navigate our environment while actively seeking and using task-relevant visual information. For this reason, an increasing number of vision researchers are employing virtual environment platforms, which offer interactive, realistic visual environments while maintaining a substantial level of experimental control. Here, we recorded eye movement behavior while subjects freely navigated through a rich, open-world virtual environment. Within this environment, subjects completed a visual search task where they were asked to find and count occurrence of specific targets among numerous distractor items. We assigned each participant into one of four target conditions: Humvees, motorcycles, aircraft, or furniture. Our results show a statistically significant relationship between gaze behavior and target objects across Target Conditions with increased visual attention toward assigned targets. Specifically, we see an increase in the number of fixations and an increase in dwell time on target relative to distractor objects. In addition, we included a divided attention task to investigate how search changed with the addition of a secondary task. With increased cognitive load, subjects slowed their speed, decreased gaze on objects, and increased the number of objects scanned in the environment. Overall, our results confirm previous findings and support that complex virtual environments can be used for active visual search experimentation, maintaining a high level of precision in the quantification of gaze information and visual attention. This study contributes to our understanding of how individuals search for information in a naturalistic (open-world) virtual environment. Likewise, our paradigm provides an intriguing look into the heterogeneity of individual behaviors when completing an un-timed visual search task while actively navigating.

Keywords: visual search, virtual environment, eye tracking, distractors, dwell time, divided attention

INTRODUCTION

Active, unconstrained visual exploration is the sensory foundation of how the majority of individuals interact with the natural world, continually seeking information from their environment. This often includes coordinated body, head, and eye movement activity. In contrast, the majority of studies that seek to understand human visual perception employ a relatively passive approach through the presentation of stimuli, whether synthetic or natural. Likewise, body, head, and even eye movements are often constrained, either explicitly or by the nature of the experimental paradigm. These factors help control the manifold sources of variability, enabling the meaningful interpretation of finite empirical data. However, as both our understanding of perception and experimental capabilities expand, an increasing number of studies have sought to explore visual processes under more natural conditions; enhancing ecological validity while maintaining construct validity (Diaz et al., 2013; Foulsham and Kingstone, 2017).

Here, we build upon a body of work that has used virtual environments to understand perceptual and cognitive processes related to visual search and navigation. Previous work investigating visual search in virtual environments have examined eye movements during object search and memory tasks (Draschkow et al., 2014; Kit et al., 2014; Li et al., 2016; Helbing et al., 2020) and have examined visual attention toward distractors (Olk et al., 2018) using timed task paradigms in traditional indoor virtual environments where items are placed in-context with surrounding environments. Previous work in the areas of spatial cognition and navigation, have created virtual maze environments to investigate visual attention during employment of allocentric and egocentric navigation strategies (Livingstone-Lee et al., 2011) and to understand the role of gender in landmark utilization (Andersen et al., 2012). In addition, virtual environments have been used to test the effectiveness of a guidance system during navigation of a train station (Schrom-Feiertag et al., 2017), and to examine spatial knowledge (Clay et al., 2019) and change detection (Karacan et al., 2010) during navigation of outdoor virtual environments. Other predecessors in this area of work have laid the ground work to integrate open sourced game engines with eye tracking to create naturalistic environments for multimodal neurophysiological research (Jangraw et al., 2014). Despite these efforts, additional work is needed to understand how visual search generalizes in a variety of real-world contexts, such as navigation and how targets are found during navigation with limited spatial (contextual) dependencies. Likewise, the capability to link gaze behavior in these complex environments, to neurophysiological processes, remains an open challenge for the field.

Measuring eye movement activity, including saccades, fixations, and blinks, has provided researchers a non-invasive way to gain valuable insight into perceptual, attentional, and cognitive processes during visual search tasks (Hoffman and Subramaniam, 1995; Kowler et al., 1995; Deubel and Schneider, 1996; Williams and Castelhano, 2019). Examining fixation metrics (e.g., number of fixations or dwell time) can indicate how individuals process visual information. For example, previous work has shown that individuals increase the number of fixations and dwell time (summation of all individual fixation durations) on informative visual objects within a scene (Loftus and Mackworth, 1978) and in the detection of changes of an object's location within a scene (Võ et al., 2010). Improved memory recall and recognition on tasks is associated with increased number of fixations (Kafkas and Montaldi, 2011; Tatler and Tatler, 2013) and increased dwell time (Hollingworth and Henderson, 2002; Draschkow et al., 2014; Helbing et al., 2020). Specifically, increased number of fixations and increased dwell time on objects during visual search tasks are linked to improved memory for those objects (Hollingworth and Henderson, 2002; Tatler and Tatler, 2013; Draschkow et al., 2014; Helbing et al., 2020). This also appears to be the case when comparing how individuals visually attend to target objects compared to distractors in the environment. Horstmann et al. (2019) found that the average number of fixations on visual targets (about 1.55) was higher compared to the average number of fixations on similar looking distractors (about 1.20) during a search task with static images. Watson et al. (2019) reported that the number of fixations on targets ranged from about 3.3–4 compared to around 2.8–3.8 fixations on distractors, during a free visual search, and reward learning task in a virtual environment. In terms of dwell time, Draschkow et al. (2014) found subjects looked about 0.6 s longer at targets as compared to distractors during visual search of static natural scenes.

Previous work suggests that visual search tasks using traditional stimuli such as static pictures may yield different findings than those incorporating real world scenarios (Kingstone et al., 2003). Research has shown notable differences in gaze metrics between simple static vs. complex dynamic visual search tasks, arguing for the increasing utilization of dynamic scenes. For instance, Smith and Mital (2013) found increased dwell time on visual objects and increased saccade amplitude during a viewing and identification task in a dynamic scene compared to a static scene. We live in a visually complex world that includes many visual points of interest, depth, motion, and contextual scene information. Therefore, real-life environments are seemingly the optimal stimuli to study naturalistic eye movement during visual search.

To this end, researchers have employed free navigation visual tasks in real-life scenarios such as walking outdoors (Foulsham et al., 2011; Davoudian and Raynham, 2012; Matthis et al., 2018; Liao et al., 2019), walking indoors (Kothari et al., 2020), driving (Land and Lee, 1994; Dukic et al., 2013; Grüner and Ansorge, 2017; Lappi et al., 2017), and shopping in a grocery store (Gidlöf et al., 2013), to name a few. Although eye tracking in a real-life scenario allows free body movement, conducting studies in real environments can be difficult if not impossible to control; every subject's unique actions makes a comparative analysis difficult. Fotios et al. (2015) noted this challenge in a study that examined eye movement for pedestrians walking down the street. Examining eye movement metrics in real life environments also limits the design of the study in terms of the availability of targets and distractors (i.e., extant objects or limited by budget) and may be limited on the ability to gather neurophysiological measures

such as electroencephalogram (EEG) recordings. Furthermore, real-world paradigms are often limited to only locally accessible environments and restrict researchers from studying more consequential scenarios where there are high demands for visual attention during a search task (e.g., looking for threat targets in a combat zone).

The use of virtual environments in perception research is an ecologically valid approach that provides the ability to conduct studies in an interactive but controlled dynamic environment (Parsons, 2015). Since eye-tracking systems can now be readily integrated with 3D rendering software (i.e., game engines), researchers can conduct eye movement studies in more realistic and immersive environments (Watson et al., 2019). Virtual environments also allow for research designs that may otherwise not be practical for a real-world implementation. For example, Karacan et al. (2010) utilized a 3D rendered virtual environment to examine shifts in gaze patterns as subjects repeatedly walked a loop path looking for isolated changes in the environment during each lap (e.g., a new object appearing, changing, and/or disappearing). The use of the virtual environment allowed for uninterrupted "physical" and visual inspection of an environment with tightly controlled visual changes. Virtual environments can accommodate research in attentional control and even allow for quantifiable interactions with objects in the scene. Helbing et al. (2020) utilized a virtual reality environment to examine memory encoding during target search of 10 different complex and naturalistic indoor rooms. Furthermore, utilizing game engines as Unity3D (Unity Technologies), can allow for the subjects to remain stationary during visual exploration of an environment and for researchers to perform synchronous acquisition of multiple physiological modalities, including respiration, electrocardiography (EKG), and EEG (Jangraw et al., 2014) that would otherwise be difficult in an ambulatory condition.

Similar to previous work in our field, the current study seeks to isolate distinct gaze behaviors associated with target objects during an active visual search of a complex environment. Here, subjects freely navigate through a virtual world while completing a self-paced visual search task identifying assigned targets placed amongst many distractors (all other objects in the virtual environment other than targets). These distractors include a wide variety of objects that are, in some cases, similar in shape, or color to the assigned target object. Additionally, objects in the world appear to be inconsistently placed along the path (e.g., laying on their side or standing upright) and randomly placed among one another with little to no scene context in regards to surrounding items or, in some cases, to the environment itself (e.g., a tuba next to a washing machine next to an airplane). This is particularly interesting since more traditional work with static scenes, has shown the importance of scene context on eye movement, memory, and search time for visual targets (Loftus and Mackworth, 1978; Henderson et al., 1999; Castelhano and Heaven, 2010; Draschkow et al., 2014). Lastly, some of the subjects are assigned a Target Condition that includes more than one particular target object in the environment. This study also includes an auditory divided attention task to increase subjects' cognitive load during a portion of the visual search task, which enables us to further investigate how subjects compensate visual attention during a self-paced task in a complex environment.

The primary aim of this study is to quantify visual search behavior during navigation of an open virtual environment and identify similarities and differences between related work that used more traditional fixed, static scenes. To this end, we quantify the difference in gaze metrics between task-relevant targets and task-irrelevant distractors (that do not provide context for locating a target) and during high and low cognitive load conditions, comparing the results to previous studies which utilized more traditional visual search and encoding paradigms. Specifically, we expect there to be an increased number of fixations and dwell time on targets, as compared to distractors (Draschkow et al., 2014; Horstmann et al., 2019; Watson et al., 2019). We likewise expect subjects will visually explore targets at a closer distance as compared to other objects in the environment. Finally, we anticipate that auditory math task will elicit changes in saccade or fixation activity, such as increased visual attention on scanned objects in the environment (Pomplun et al., 2001; King, 2009; Buettner, 2013; Zagermann et al., 2018) or a change in exploratory behavior (i.e., reduction in speed and number of objects viewed) of the environment, due to increased cognitive load.

MATERIALS AND METHODS

Subjects

Forty-Five subjects, recruited from the Los Angeles area, participated in this study [17 females with mean age ± standard deviation (SD) = 36.8 ± 12.3 years, 28 males with mean age ± SD = 41.6 ± 14.4 years]. All subjects were at least 18 years of age or older and able to speak, read, and write English. All subjects signed an Institutional Review Board approved informed consent form prior to participation (ARL 19–122) and were compensated for their time. All subjects had normal hearing and normal or corrected-to-normal vision and had normal color vision. All subjects completed a web-based pre-screen questionnaire containing eligibility, demographic, and game-use questions. Additional color vision and visual acuity screening was conducted in-lab to ensure a minimum of 20/40 vision, using a standard Snellen Chart, and normal color vision, assessed with a 14-plate Ishihara color test. Any subject who did not pass the screening process was not included in the study. Subjects completed a simulator sickness screening questionnaire, the Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993), before and after Pre-Test training (see below) and then again after the main study task. The mean and SD for the Total Weighted Score from the SSQ was 12.6 ± 16.4 before the system training task, 32.1 ± 30.7 after the training task, and 38.1 ± 37.5 after the main study task. As part of the questionnaire, subjects answered questions relating to video game experiences and weekly usage of video games. The average number of years playing video games was 28.0 ± 11.4 years. The mean age when subjects began playing video games was 12.8 ± 7.1 years. Over half of subjects (51%) reported playing video games <2 h a week. Almost a third (29%) of the subjects reported playing video games 2–7 h a week. The remaining 20% of subjects reported playing video games for >8 h

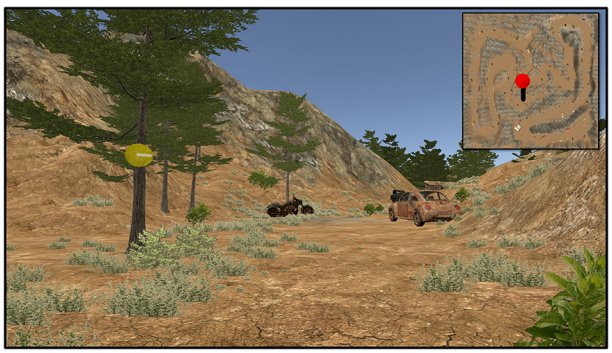


FIGURE 1 | First person point of view near the beginning of the task. Trail makers (yellow circles with direction indicator) were placed on trees throughout the environment. Targets were assigned to each Target Condition to count: Humvee, motorcycle (shown), aircraft, and furniture. Distractors were any object in the environment not assigned to the subject (e.g., tires, dumpster, Humvees for anyone not in the Humvee Condition). Inset (not visible during experiment) shows the current position on the complete map.

a week. Six subjects were later removed from the main study analysis ($N = 39$) and an additional one subject was not included in the Math Task analysis ($N = 38$) due to reasons detailed below.

Procedure

During the experimental session, subjects participated in four separate tasks: a go/no-go serial visual presentation, an old/new recognition task, and two virtual environment tasks. However, only results from the virtual environment training and navigation tasks are described here. The stimuli in the other tasks were unrelated to the virtual environment.

Overview

Subjects were asked to freely navigate the virtual environment with the goal of searching for and counting their assigned target objects. All subjects were randomly assigned to one of four Target Conditions: Humvee Condition ($N = 15$ subjects), Motorcycle Condition ($N = 14$ subjects), Aircraft Condition ($N = 9$ subjects), or Furniture Condition ($N = 7$ subjects). The Aircraft and Furniture Conditions were introduced later in the data collection, which was eventually halted due to restrictions on in-person studies, hence the lower subject numbers. The aircraft and furniture targets were already present in the environment prior to introduction of the two new Conditions, thus, all subjects in every Target Condition navigated the same environment with the same objects in the same order (**Figure 1**). Natural landscape features and trail markers provided a suggested path through the virtual environment (although subjects could freely explore in any chosen direction).

System Training Task

A training task was used to acclimate subjects to navigation in the virtual environment via the keyboard and mouse. Movement was controlled with the W/A/S/D keys: "W" moved the subject in the forward direction, "A" allowed the subject to move left, "S" moved the subject backwards, and "D" allowed the subject to move right. A computer mouse was used to control the camera orientation or viewport (i.e., first person perspective) while in the virtual environment. This training environment was similar to the virtual environment used during the main task but contained different objects. This training task also ensured subjects were not acutely susceptible to simulator sickness.

Testing Setup

The experimental setup for this study combined multiple physiological modalities: eye-tracking, EEG, electrocardiography (EKG). Here, we described the relationship between task features, performance, and eye movement behavior. Other modalities, such as EEG and EKG, will be discussed in future reports and are not included in the current study.

All tasks were run using custom software built in the Unity 3D environment (Unity Technologies) run on the standard Tobii Pro Spectrum monitor (EIZO FlexScan EV2451) with a resolution of 1,920 x 1,080 pixels. Subjects were seated at a distance of ∼70 cm from the monitor. Eye tracking data were collected with a Tobii Pro Spectrum (300 Hz). In addition to obtaining gaze position and pupil size, the Tobii Pro SDK was used to calculate the 3D gaze vector (invisible ray representing the instantaneous gaze direction) and identify the gaze vector collision object (first object in the Unity environment that collides with the 3D gaze vector) for each valid sample. The eye tracking data were synchronized with the game state, keyboard, mouse, and EEG data using the Lab Streaming Layer protocol (Kothe, 2014). A standard 5-point calibration protocol was used to calibrate the eye tracker. The Tobii Pro Spectrum has an average binocular accuracy of 0.3°, binocular precision (root mean square) of 0.07°, and detects 98.8% of gazes (Tobii Pro, 2018). However, no verification of these error metrics was performed for this study. Head movement was not restricted in terms of head support or a chin rest. However, subjects were asked to maintain an upright, yet comfortable posture to minimize large upper body movements and maintain proper alignment with the eye tracker.

Virtual Environment Description

Targets were placed in a random sequence at semi-regular intervals along the path in the virtual environment. The location of all targets and objects in the environment were the same for all subjects. A general layout of the environment, indicating all target locations, is shown in **Figure 2** and target examples are shown in **Figure 3**. As stated previously, all objects were embedded in the environment in such a way that each appeared randomly placed with no context gained from neighboring objects. Thus, targets (and distractors) appeared along the path and could not be anticipated by surrounding objects that may give the subjects an indication that a target was visually missed, present, or forthcoming. Subjects all started at the same point on the virtual environment. Trail markers ($N = 19$) were placed along the trail for general navigational guidance. There were 15 targets total for each Target Condition. The same model of Humvee was used for all the *Humvee targets* and the same model of motorcycle was used for the *motorcycle targets*. For the *aircraft targets*, models varied and included helicopters, bi-planes, and one glider. For the *furniture targets*, objects included variations

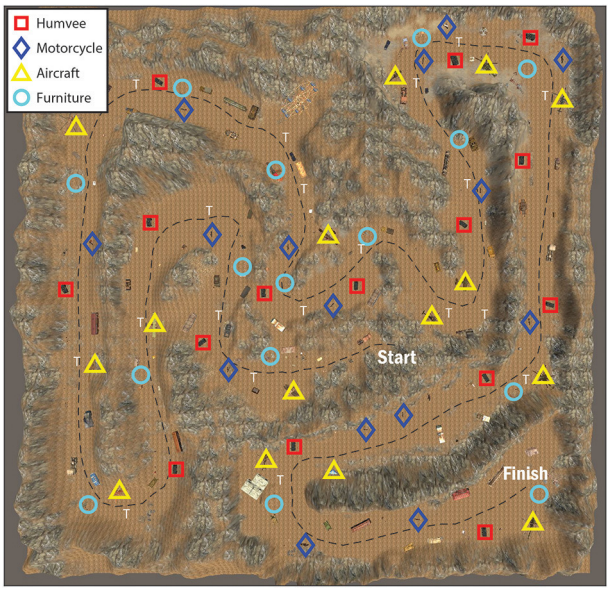


FIGURE 2 | General layout of the virtual environment map. The black checkered line represents an example subject's path from the starting area to the finish. Target icons are as follows: furniture (light blue circle), motorcycle (dark blue diamond), aircraft (yellow triangle), and Humvee (red square). Trail markers are present throughout the path and are indicated by a white "T."

FIGURE 3 | Example pictures of the targets for each Target Condition as they appear in the virtual environment. Motorcycles and Humvees (top row) did not vary in model but did vary in how they were positioned on the trail (i.e., motorcycle against a tree or at an angle by a rock). Both aircrafts (middle row) and furniture (third row) targets varied in shape, size, and positioning on the trail.

such as beds, grandfather clocks, tables, and a variety of seating furniture (e.g., sofa, dining chair). Sizes varied for the furniture with the chair being the smallest and the bed being the largest furniture target. Around 166 additional objects were included in the virtual environment that were not an assigned target to any Target Condition. These additional objects included, but were not limited to, cars, trucks, tanks, an oven, a drum set, a Ferris wheel, a pile of tires, dumpsters, and shipping containers. For analysis, a distractor was defined as any visual object in the environment not belonging to the specified Target Condition and included objects assigned as targets to other Target Conditions (e.g., Humvees were considered distractors for the Motorcycle Condition). Terrain (e.g., trees, hillside, grass, path) and the sky were not included in the analysis unless explicitly mentioned.

Subject Instruction and Navigation

Subjects were instructed to search and count (mentally) when they saw a target assigned to their Target Condition. Subjects were encouraged to stay on or near the trail (and at times were verbally reminded by research staff) to make sure they encountered all objects, but were free to navigate as desired. Midway (8 min) into the session an auditory Math Task (divided attention task) was administered (see below for details). Subjects had up to 20 min to progress through the virtual environment and reach the finish. If subjects did not complete the task in 20 min, and if they did not encounter (as determined by their gaze vectors) at least 10 targets in the virtual environment, then their data was removed from statistical analysis. For this reason, data from two people in the Furniture Condition were removed from all analysis. In addition, one subject in the Humvee Condition reported feeling unwell during testing and experienced difficulty in navigating the environment (i.e., did not follow the path) and thus, their data was also removed from the analysis. The average time to complete navigation of the virtual environment was about 12 (± 2) min. After completion, subjects were asked to recall how many targets they saw during the navigation task.

Additional Math Task

Starting at the 8 min mark, an auditory math problem was presented to the subjects. An auditory recording of a set of 3 to 4 numbers, with values between 0 and 9, was played for the subject through headphones (e.g., "4," pause, "2," pause, "8," tone, subject reports "14"). A pause of 3–4 s separated each number in a set, and each set was followed by a tone. After the tone, subjects verbally reported the sum of numbers aloud to the experimenter. During the Math Task, subjects were instructed to continue navigating through the virtual environment and continue searching and mentally counting their targets. This Math Task was repeated two more times (with different sets of numbers), for a total of three summation responses. There was an 8–30 s break between each set of numbers. Because the primary search task was self-paced, it is possible that a subject would finish exploring the virtual environment (reach the end of the path) without completing the Math Task. Only one subject (in the Humvee Condition) did not complete the Math Task prior to finishing the navigation task and for this reason, their data was removed from the Math Task analysis.

Data Extraction and Analysis

Fixation Detection and Object Labeling

Blinks were identified from stereotyped gaps in the gaze position data (Holmqvist et al., 2011) while saccades (and corresponding fixations) were detected using a standard velocity-based algorithm (Engbert and Kliegl, 2003; Engbert and Mergenthaler, 2006; Dimigen et al., 2011) adapted from the EYE-EEG plugin (http://www2.hu-berlin.de/eyetracking-eeg).

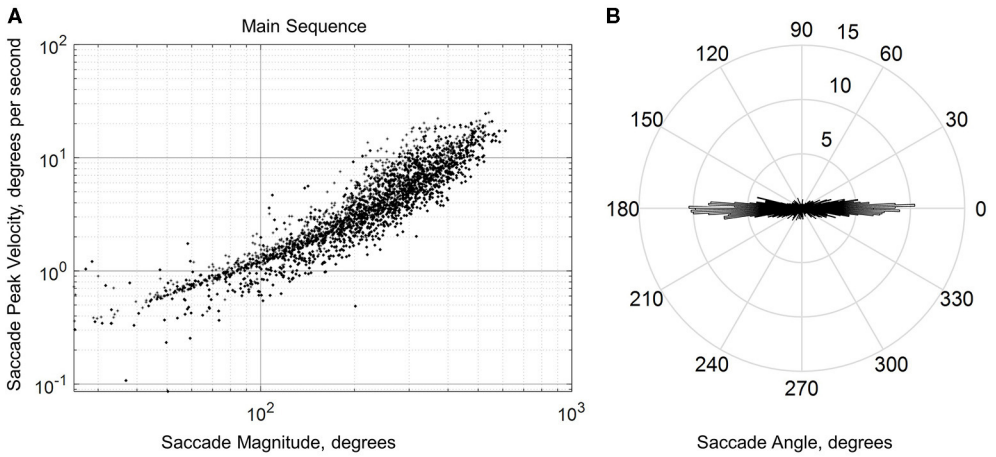


FIGURE 4 | Saccade main sequence (in log-log coordinates) **(A)** and angle distribution **(B)** from a sample of the first 500 saccades from five representative subjects. The saccade angle distribution histogram has a one-degree resolution, with the radial axis showing average number and the angular axis showing average angle across all five subjects.

Specifically, velocity thresholds for saccade detection were based on the median of the velocity time series, smoothed over a 5-sample window, for each subject. Thresholds were computed independently for horizontal and vertical components. In this study, we used a velocity factor of six (six times the median velocity), a minimum saccade duration of 12 ms, and a minimum fixation duration (i.e., inter-saccadic interval) of 50 ms. We kept only the largest saccade and subsequent fixation if two or more saccades were detected within the minimum fixation duration window. Visual inspection of the saccades shows the expected relationship between saccade peak velocity and saccade magnitude (i.e., main sequence; **Figure 4A**). Only the first 500 saccades of 5 subjects are shown in this figure due to the large amount of saccades generated by each subject (∼8–20 min of eye tracking). These 5 subjects were chosen randomly and are representative of the entire subject data set. These saccade distributions excluded blinks, dropouts, saccades with a duration shorter than 12 and >100 ms, and peak velocities outside of a range of 25 and 1,200 degrees per second. The distribution of the saccade angle showed a strong tendency for subjects to scan the horizon (**Figure 4B**). Finally, blinks were defined as gaps with a duration ranging from 50 to 500 ms and dropouts were defined as any gap with a duration <50 or >500 ms (any gap not considered a blink). While the detection algorithm used in this study was developed for ballistic saccades, visual inspection revealed that it was reasonably successful at separating saccades from other eye movement features such as smooth pursuit and optokinetic responses (**Figure 5A**).

After initial saccade detection, fixations of <100 ms were discarded and not used in any subsequent analysis (Ouerhani et al., 2004; Mueller et al., 2008; Andersen et al., 2012). In addition to standard metrics associated with fixations (e.g., duration), each fixation was assigned a virtual environment object label using the following approach. Every valid gaze sample returned a corresponding object that was the result of the gaze vector collision. The object with the highest percentage of collisions over the fixation epoch was assigned as the "fixation object" (**Figure 5A**). Target fixations were labeled as such if the highest percentage of collisions were on a target (e.g., motorcycles for the Motorcycle Condition) and this number amounted to at least 10% of all gaze samples for that fixation. Distractor fixations were labeled using the same metric; receiving a distractor designation label if at least 10% of the gaze samples included the same distractor object. This 10% threshold was utilized to identify the primary fixation object and reduce the chance of including adjacent or background objects. Objects could be erroneously included in a fixation epoch from either gaze vector estimation error or having a relative position directly behind the primary fixation object. We selected the 10% value by assessing fixation labels at a range of thresholds: 0, 1, 10, 20, and 50% (**Figures 5B,C**). The 10% threshold appeared to be a middle ground between reducing the chance of erroneous fixations without removing a large number of meaningful fixations. Fixation data from three subjects (two from the Aircraft Condition and one from the Humvee Condition) had a high dropout rate (high number of invalid samples). Thus, these three subjects were removed from the analysis.

Calculation of Fixation Variables for the Main Study Analysis

From the fixation data for the targets, the following variables in **Table 1**, were calculated. The Self-Reported Target Count and the Gaze-Validated Target Count (the instances where the ray cast at the fixation intersected with the object using the above fixation labeling approach) were calculated to compare subjective inventory with detected target fixations. To identify if our approach was sensitive enough to detect increased visual attention on targets, the Mean Number of Fixations, Mean Dwell Time, and Mean Distance were compared. Distance is included to provide a relative measure of how "close" subjects approached

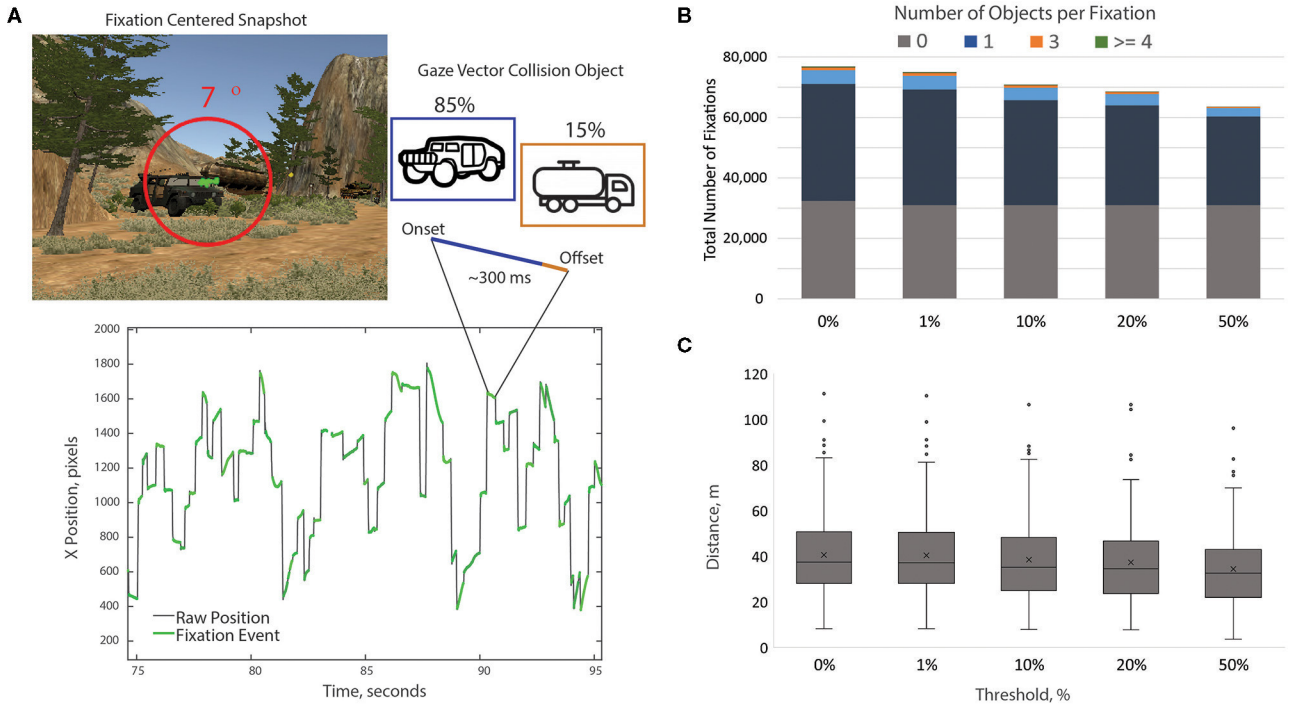


FIGURE 5 | Fixation labeling approach. Saccades and fixations are detected from the raw X/Y gaze position time series. Note, that the inter-saccadic intervals (fixations) exhibit a range of velocities and include stable fixations, smooth pursuit movements, and optokinetic responses. However, these velocities were significantly lower than the ballistic threshold of the saccade detection algorithm. Each valid gaze sample is associated with an object via gaze vector collision. Fixations are then associated with objects if the dominant object (excluding terrain and sky) is contained in at least 10% of the samples of the epoch. **(A)** Alternative thresholds were examined and the 10% threshold appeared to filter out erroneous labels without excluding large amounts of meaningful data. The impact of the threshold value on the number of **(B)** and distance to **(C)** objects with all fixation epochs.

objects in the environment. Importantly, although the units here are given in meters, we acknowledge that this metric is not an equivalent analog to the real world (i.e., meters in the virtual environment may not reflect an actual meter in real life). Variation in object size and structural diversity impacted these particular fixation metrics. For instance, object surface area in the virtual environment was shown to be a large covariate with Mean Number of Fixations (Spearman's rho = 0.719, $p = 0.000$), Mean Dwell Time (Spearman's rho = 0.630, $p = 0.000$), and Mean Distance (Spearman's rho = 0.558, $p = 0.000$). The larger the object, the increased chance a subject has to see it at any given viewing point, regardless of attentional focus. To help account for this bias, three additional variables, Normalized Number of Fixations, Normalized Dwell Time, and Normalized Distance were calculated utilizing the Global Number of Fixations, Global Dwell Time, and Global Distance variables. The global values were then used to normalize the means associated with the diversity of Target for each Target Condition. Because of the large size disparity between objects in the virtual environment, normalization by dividing gaze data by object size (utilizing either the 3D volume or 2D profile) resulted in a large bias toward the smaller targets.

As an additional analysis, we included a comparison of gaze data between just the Humvee and Motorcycle Conditions to identify differences in gaze behavior between the Humvee and motorcycle objects. This analysis provided evidence of how subjects in two different Target Conditions examined these two particular objects differently and how target assignment impacted gaze metrics. The Humvee and the Motorcycle Conditions were utilized in this way because these two conditions were comparable in subject numbers ($N = 13$ and 14, respectively) and target attributes (i.e., same object model throughout the environment). In addition, these two conditions had targets that differed greatly in size and, as previously stated, we expected there to be differences in non-normalized gaze metrics, simply due to size of the object alone.

For the Math Task, fixation data from the following two time periods was compared: outside (before and after) and during the Math Task. To see if subjects changed the rate at which they fixated objects due to the Math Task, the Mean Duration of Individual Fixations *on objects* and Fixation Rate were compared between these time periods. To determine if subjects compensated for divided attention during the Math Task by reducing the overall amount of visual attention devoted to each object, the Mean Number of Fixations *per each object* and Mean Dwell Time *per each object* was compared between the two time periods. Object Rate, the number of distinct objects that were fixated per unit time, was compared across the two time

TABLE 1 | Dependent and independent variable list and definitions.

Variable	Definition
Target condition	Each subject was randomly assigned to one of four groups, named for the target they were assigned (i.e., Humvee Group assigned to look for Humvee targets)
Self-reported target count	Total count of targets that the subject reported seeing during exploration of the virtual environment
Gaze-validated target count	Total number of targets having at least one qualifying fixation associated with that target
Mean number of fixations	Average number of qualifying fixations per each object
Mean dwell time	Average total duration of fixations per each object
Mean distance	Average distance from the object when each associated fixation occurred
Global number of fixations	Average number of fixations for that particular object across all subjects
Global dwell time	Average dwell time for that particular object across all subjects
Global distance	Average distance from where that object was fixated across all subjects
Normalized number of fixations	Mean Number of Fixations subtracted from the Global Number of Fixations
Normalized dwell time	Mean Dwell Time subtracted from the Global Dwell Time
Normalized distance	Mean Distance subtracted from the Global Distance
Mean duration of individual fixations	Average of all individual fixations across all Target Conditions and objects (targets and distractors) in the environment
Fixation rate	Summation of all fixations during the Math Task (or outside of the Math Task) divided by the total time spent in that time period
Object rate	Total number of distinct objects that were fixated per unit time
Blink rate	Summation of all blinks during the Math Task (or outside of the Math Task) divided by the total time spent in that time period
Proportion of fixations on objects	Summation of fixations on objects (as opposed to terrain or sky) divided by sum of all fixations overall
Position velocity	Average change in position over time

periods to capture differences in visual scanning behavior. To understand if subjects compensated for divided attention during the Math Task by reducing visual attention on particular objects and instead focused on background scenery, the Proportion of Fixations on Objects was compared between the two time periods. To examine if subjects speed up or slowed down their navigating through the environment, the Position Velocity was compared between the two time periods. Lastly, Blink Rate examined if subjects changed the number of blinks per unit of time with increased cognitive load.

Statistical Analysis

To summarize, data from three subjects were removed due a high dropout rate, data from two subjects were removed due to not encountering the minimum threshold of targets, and data from one subject (who reported feeling ill during the testing) had navigational issues was removed, bringing the final inclusion of $N = 39$ subjects for analysis. In addition, one subject finished navigating the environment prior to the completion of the Math Task, for a total of $N = 38$ for that analysis. For the remaining subjects', a normal distribution was assessed for all fixation variables using Kolmogorov-Smirnov and Shapiro-Wilk tests for normality. Parametric tests (i.e., Paired Samples *t*-test, MANOVA) were used for variables with normal distributions and non-parametric tests (i.e., Related-Samples Wilcoxon Signed Rank Test, Friedman Test) for non-normal distributions. For this reason, non-parametric statistical methods were utilized for the measures of Self-Reported Target Count, the Gaze-Validated Target Count, Blink Rate, and Position Velocity. All other variables had a normal distribution and parametric tests were used for comparative analysis. Outliers in the data were designated as samples/observations that were greater or less than three standard deviations from the mean. Outliers were removed from the data prior to analysis and includes one person's data for Mean Distance and Normalized Dwell Time ($N = 38$ for analysis with these measures) and one person's data for Mean Duration of Individual Fixations *on objects*, Mean Dwell Time *per object,* and Blink Rate during the Math Task analysis ($N = 37$ for analysis with these measures). In addition, Self-Reported Target Count was missing for six additional individuals (who did not report an answer when prompted) and one outlier was removed from the Self-Reported Target Count for a total of $N = 32$ for analysis with this measure. A $p < 0.05$ was considered significant for all analyses and all analysis was conducted with IBM SPSS Statistics for Windows (Version 22, Armonk, NY: IBM Corp, Released 2013) software.

RESULTS

Confirmation of Fixated Targets

On average, subjects reported the correct number of targets observed in the environment. A Related-Samples Wilcoxon Signed Rank Test compared the Self-Reported Target Count and the Gaze-Validated Target Count. There was no statistical difference between the two counts of the targets by subjects or identified by the system ($Z = -0.573$, $p = 0.567$). Median target counts were 15 for the Self-Reported and 14 for the Gaze-Validated.

General Eye-Gaze Measurement Outcomes

On average, individual fixations had a median duration of about 0.30 s (300 ms) and a mean Fixation Rate of ∼2.06 fixations-per-second throughout the main task when short fixations were removed. If short fixations were included (removal of the 100 ms cut-off threshold), the median duration decreases to 0.29 seconds (∼4%). Therefore, we have determined that the removal of those fixations with a duration of less than 100 ms has a minimal effect on the individual duration of fixations outcome. Subjects looked at objects (e.g., motorcycle, dumpster, trail markers) in the virtual environment, with a Mean Number of Fixations of 7.1 and for a Mean Dwell Time of 2.60 s per each object. We found that fixations on the surrounding terrain and sky comprised, on average, about 47% of all fixations.

Two separate one-way multivariate analysis of variances (MANOVAs) determined the effect of Fixation Object (target or distractor) on the normalized and non-normalized Mean Number of Fixations and Mean Dwell Time. There was a significant effect of Fixation Object for both non-normalized [$F_{(2,37)} = 23.84$, $p = 0.000$] and normalized gaze data [$F_{(2,36)} = 22.54$, $p = 0.000$; **Figures 6A–C, D–F**]. Two separate Univariate analysis of variances (ANOVAs) examined how Mean Number of Fixations and Mean Dwell Time differed depending on Fixation Object. Subjects significantly increased both the Mean Number of Fixations [$F_{(1,38)} = 35.73$, $p = 0.000$] and the Mean Dwell Time [$F_{(1,38)} = 48.84$, $p = 0.000$] for targets compared to distractors (**Figures 6A,B**). Two additional Univariate ANOVAs showed that Normalized Number of Fixations [$F_{(1,37)} = 44.48$, $p = 0.000$] and Normalized Dwell Time [$F_{(1,37)} = 42.54$, $p = 0.000$] also increased significantly for targets compared to distractors (**Figures 6D,E**). Mean Distance was compared between Fixation Objects using a Univariate ANOVA (**Figure 6C**). Subjects were significantly closer (less distance) to fixated targets compared to fixated distractors in the virtual environment [$F_{(1,37)} = 12.99$, $p = 0.001$]. A separate Univariate ANOVA showed that Normalized Distance [$F_{(1,38)} = 18.53$, $p = 0.000$] was also significantly less, on average, for targets (**Figure 6F**).

Fixation Differences for the Humvee Condition and Motorcycle Condition

Gaze data from the Humvee Condition and the Motorcycle Condition was used to quantify differences in the fixation metrics for the two target objects that were similar in terms of number, consistency, and dispersion along the path. Two separate two-way MANOVAs determined the effect of Target Condition and Fixation Object (limited to Humvees and motorcycles for this analysis) and the interaction of Target Condition and Fixation Object on the normalized and non-normalized Mean Number of Fixations and Mean Dwell Time. Overall, both non-normalized and normalized Mean Number of Fixations and Mean Dwell Time were significantly dependent upon the main effect of Fixation Object and the interaction between Target Condition and Fixation Object (**Figure 7** and **Table 2**). Four separate Univariate ANOVAs determined that for the main effect of Fixation Object, only the Mean Number of Fixations (non-normalized) was significantly higher overall for the Humvee object compared to the motorcycle (**Table 2**). Four other separate Univariate ANOVAs determined that the interaction between Target Condition and Fixation Object was significantly different for non-normalized and normalized variables (**Table 3**). We found a significant main effect of Fixation Object for the Mean Number of Fixations, where there was an overall greater number of fixations for the Humvees compared to motorcycles (**Table 3**). Fixation Object was not a significant main effect for Mean Dwell Time, Normalized Number of Fixations, or Normalized Dwell Time. Tukey *Post-hoc* determined significant differences in those interactions. Both Target Conditions had significantly greater Mean Number of Fixations and greater Mean Dwell Time devoted to their targets, compared to the other object ($p < 0.01$, Tukey *Post-hoc*). Both Target Conditions had increased Mean Number of Fixations and Mean Dwell Time on their respective targets compared to that object for the other Target Condition (i.e., the Humvee Condition focused on the Humvees significantly more than the Motorcycle Condition focused on Humvees) ($p < 0.01$, Tukey *Post-hoc*). For these comparisons, the same pattern of *Post-hoc* analysis statistical significance was found for the Normalized Mean Number of Fixations and Normalized Dwell Time ($p < 0.05$). The Mean Number of Fixations for the Motorcycle Condition was significantly greater for the Humvee target compared to the Mean Number of Fixations for the Humvee Condition and the motorcycle target ($p < 0.05$, Tukey *Post-hoc*). No other differences were statistically significant.

Effect of Cognitive Load

The total time spent on the Math Task was ∼150 s (∼2.5 min), compared to the time spent outside of the Math Task (before and after) 713 s (∼12 min). Paired samples *t*-test determined that subjects did not significantly change the Mean Duration of Individual Fixations *on objects* ($t_{36} = 0.03$, $p = 0.979$) during the Math Task compared to outside of the Math Task. However, Paired samples *t*-tests showed that subjects significantly decreased their Fixation Rate ($t_{37} = -2.91$, $p = 0.006$; **Figure 8C**). This discrepancy was explained by the relative increase in Blink Rate during the Math Task (Related-Sample Wilcoxon Signed Rank Test, $Z = 3.78$, $p = 0.000$; **Figure 8E**). There was also a significant reduction in the Mean Number of Fixations ($t_{37} = -5.67$, $p = 0.000$) *per object* and the Mean Dwell Time ($t_{36} = -4.51$, $p = 0.000$) *per object* during the Math Task, as compared to outside the Math Task (**Figures 8A,B**). In contrast, Object Rate increased significantly during the Math Task compared to outside the Math Task ($t_{37} = 3.44$, $p = 0.001$; **Figure 8D**). Interestingly, a Paired samples *t*-test showed that the Proportion of Fixations on Objects in the virtual environment (as opposed to fixations on terrain or sky) did not significantly change during the Math Task portion compared to outside of the Math Task ($t_{37} = 0.16$, $p = 0.873$). Additionally, a Related-Sample Wilcoxon Signed Rank Test showed that subjects significantly reduced their Position velocity, the speed at which they progressed through the environment, during the Math Task compared to outside the Math Task ($Z = -4.87$, $p = 0.000$; **Figure 8F**).

DISCUSSION

In this study, we demonstrate how an open-world, virtual environment can be used to identify task-relevant gaze behavior during navigation. Our approach enables us to collect meaningful, object-centered, gaze information during visual search in a cluttered landscape without restricting virtual head movement (i.e., camera position and orientation). Consistent with previous studies, we show a clear distinction in gaze behavior between target and distractor objects. Moreover, we quantify how this gaze behavior changes when subjects' attention is divided between visual search and secondary auditory task. Our results build on previous work using virtual environments (Karacan et al., 2010; Livingstone-Lee et al., 2011; Andersen et al.,

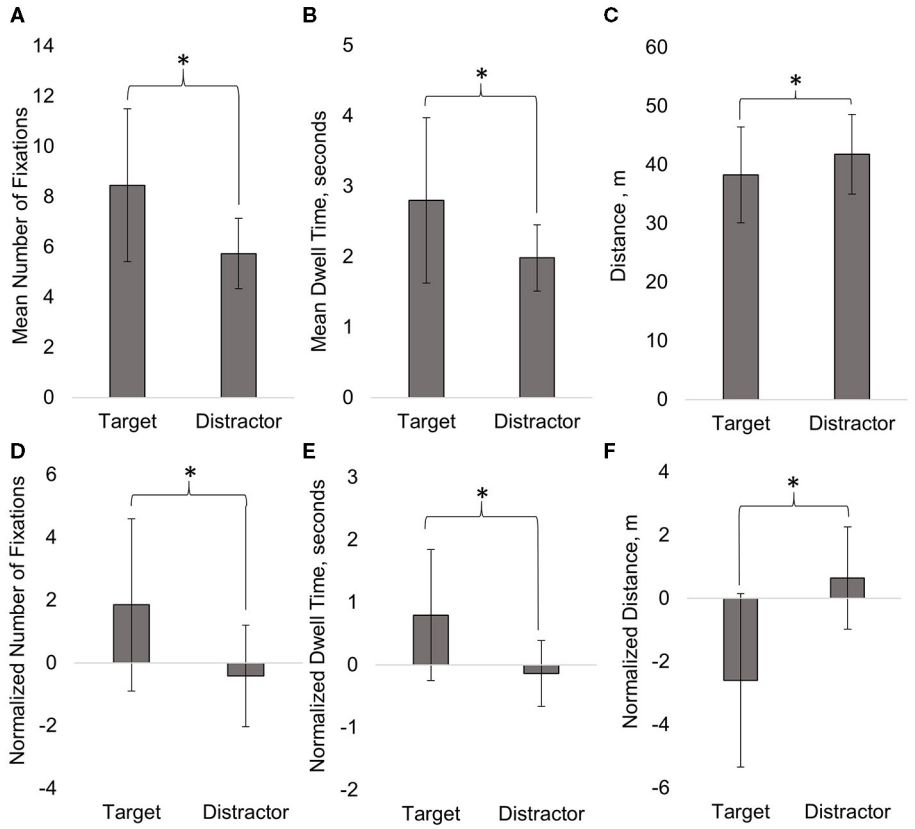


FIGURE 6 | The Mean Number of Fixations **(A)**, Mean Dwell Time **(B)**, Mean Distance **(C)**, Normalized Number of Fixations **(D)**, Normalized Dwell Time **(E)**, and Normalized Distance **(F)** were significantly greater for targets compared to distractors. Mean ± SD (error bars) are shown on graph. *$p<0.05$.

2012; Draschkow et al., 2014; Jangraw et al., 2014; Kit et al., 2014; Li et al., 2016; Schrom-Feiertag et al., 2017; Olk et al., 2018; Clay et al., 2019; Helbing et al., 2020), extending the search space and incorporating a secondary task while maintaining the temporal and spatial precision needed for neurophysiological analysis.

General Discussion

Here, we observed a direct link between gaze activity and specific objects within the virtual environment. Overall, subjects looked at targets significantly more often and longer than distractors. This confirms our initial hypothesis, based on previous studies in more restricted experimental contexts, and demonstrates the feasibility of gaze analysis in dynamic (constantly changing) environments. This study is unique in that we acquired data from a relatively large number of subjects ($N = 39$) navigating a detailed and complex virtual environment, but were still able to identify distinct condition-level gaze dynamics. This fixation and object-level precision enables meaningful inferences from the concurrent use of EEG (not reported here).

Comparable General Outcomes to Literature

Overall, the basic eye movement outcomes were comparable to those found in literature with a few notable differences. We found that individual fixations had a duration of about 320 ms with a Fixation Rate of 2.06 fixations-per-second. This is comparable to a Foulsham et al. (2011), who found an average of 2 fixations-per-second and an individual fixation duration of 441 ms for subjects who watched a video of path they previously navigated. Subjects looked at objects (including targets and distractors) in the virtual environment, on average, about 7.1 times, comparable but higher than the number of Mean Number of Fixations reported by Zelinksy (2008) who found an average of 4.8 fixations to detect and locate military tanks in a realistic scene. In terms of dwell time, Clay et al. (2019) reported a Mean Dwell Time of 5.53 s on visual objects for subjects who freely navigated and observed houses (large objects) in a virtual town. This was higher than our reported Mean Dwell Time of 2.60 s per each object, perhaps due to the fact that the objects in our world were considerably smaller on average than the houses in the Clay et al. (2019) study and in that study subjects were freely exploring without searching for specific targets. Finally, ∼47% of all our fixations were on the sky or terrain (path) surrounding the environment, which is comparably within range of what others have also noted on visual attention when walking. Foulsham et al. (2011) found that about 29% of fixations focused on the path ahead of where subjects were walking and Davoudian and Raynham (2012) who found about 50% of fixations were focused on the walking path. Overall, our

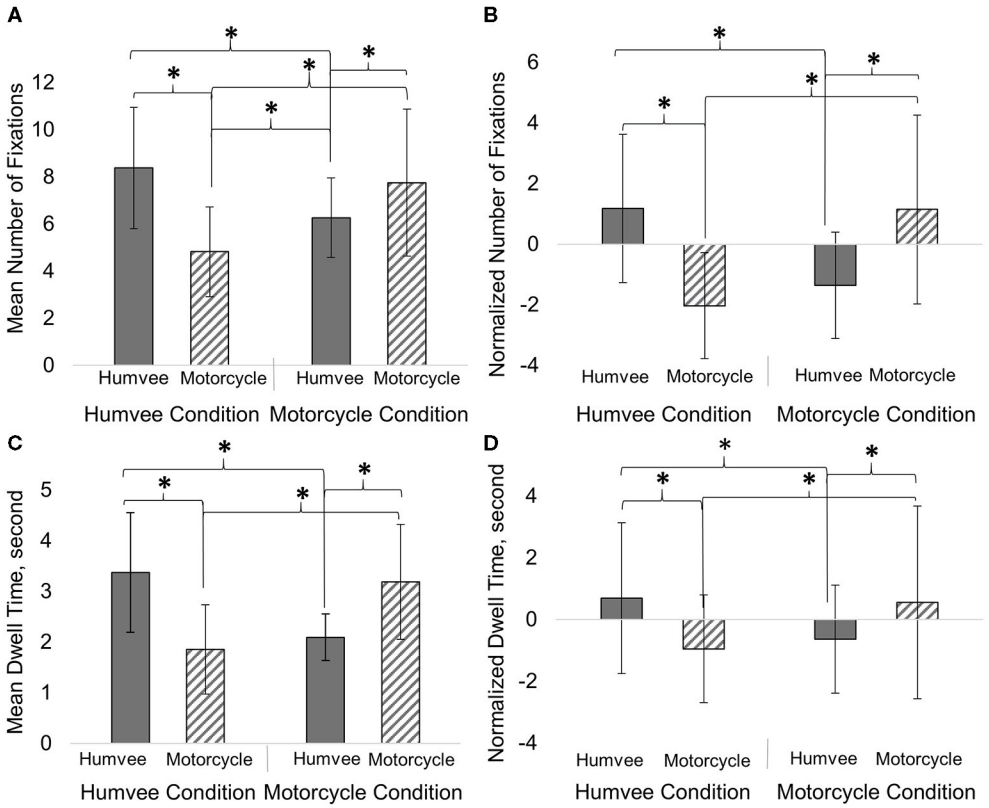


FIGURE 7 | Both Target Conditions significantly increased Mean Number of Fixations **(A)**, Normalized Mean Number of Fixations **(B)**, Mean Dwell Time **(C)**, and Normalized Dwell Time **(D)** for their respective targets. Subject's increased visual attention toward their respective targets compared to the Fixation Object (a distractor). Mean ± SD (error bars) are shown on graph. *$p<0.05$.

general outcomes showed reasonable comparison to those found in previous work.

Increased Number of Fixations and Dwell Time on Targets Compared to Distractors

In our study, we found subjects increased their visual attention (as measured by Mean Number of Fixations and Mean Dwell Time) on targets as compared to distractors. The Number of Fixations was significantly greater for targets compared to distractors. This 47% increase in fixations on targets over distractors was comparable to previous work. In a traditional visual search task with static images, Horstmann et al. (2019) found an approximate increase of 33% in Mean Number of Fixations for targets compared to target-similar distractors. Watson et al. (2019) found approximately a 10% increase overall in fixations for targets compared distractors in a study using a reward learning visual search task. We observed 41% increase in Mean Dwell Time for targets compared to distractors. This outcome was comparable to work by Draschkow et al. (2014) who observed around a 30% increase in Mean Dwell Time on targets compared to distractors during a timed visual search task of complex static naturalistic scenes. Our result, showing increased overt visual attention on targets, supports our claim that subtle changes in visual search behavior can be quantified in complex and dynamic virtual environments. Overall, our results were in line with previous studies, supporting the validity to our approach and processing methods.

It should be noted that the overall Mean Number of Fixations for both targets and distractors reported here, is greater than what has generally been found in many of the previous studies. This could be due to the task design and nature of the environment and task. Even though subjects were given a maximum time of 20 min to complete the task, subjects were not instructed to find their targets as quickly as possible, as is the case in many visual search studies. Thus, subjects had more time to visually inspect all objects in the environment, without feeling rushed. Our environment contained 15 targets for each condition and ∼211 distractors (including other Target Conditions' targets, a ratio of targets to distractors of about 1:14). Increasing the number of search items or the number of distractors can impact the working memory load and reduce visual search efficiency (Palmer, 1995; Wolfe, 2007, 2012; Zelinsky, 2008; Gidlöf et al., 2013), especially in complex naturalistic environments (Wolfe, 1994a; Gidlöf et al., 2013). Therefore, the increased number of fixations observed in our study could be due to the subject's self-pace progression through the environment and the particular target to distractor ratio. Alternatively, movement through virtual environments generate a more diverse set of eye movements (e.g.,

TABLE 2 | MANOVA for the non-normalized and normalized gaze data.

	Wilks lambda F, $df_{(2,24)}$	*P*-value	Effect size, η_p^2
Non-normalized gaze data (Mean number of fixations and mean dwell time)			
Target condition	0.08	0.928	0.006
Fixation object	6.68	0.005*	0.357
Target condition × fixation object	19.64	0.000**	0.621
Normalized gaze data (Normalized number of fixations and normalized dwell time)			
Target condition	0.07	0.93	0.006
Fixation object	0.47	0.632	0.038
Target condition × fixation object	20.6	0.000**	0.632

$^{**}p < 0.01$, $^{*}p < 0.05$.

TABLE 3 | Univariate ANOVAs for the non-normalized and normalized gaze data.

	Wilks lambda F, $df_{(1,25)}$	*P*-value	Effect size, η_p^2
Non-normalized gaze data (Mean number of fixations and mean dwell time)			
Mean number of fixations			
Target condition	0.11	0.748	0.004
Fixation object	7.49	0.011*	0.22
Target condition × fixation object	38.56	0.000**	0.607
Mean dwell time			
Target condition	0.02	0.902	0.001
Fixation object	0.9	0.352	0.035
Target condition × fixation object	36.75	0.000**	0.595
Normalized gaze data (Normalized number of fixations and normalized dwell time)			
Mean number of fixations			
Target condition	0.14	0.707	0.006
Fixation object	0.62	0.438	0.024
Target condition × fixation object	41.38	0.000**	0.623
Mean dwell time			
Target condition	0.07	0.796	0.003
Fixation object	0.96	0.336	0.037
Target condition × fixation object	38.19	0.000**	0.604

$^{**}p < 0.01$, $^{*}p < 0.05$.

smooth pursuit and optokinetic responses) which can impact the detection and labeling of ballistic saccades and inter-saccadic intervals (i.e., fixations).

Additionally, for some of our Target Conditions, target characteristics could have led to an overall high mean Number of Fixations on targets and distractors. For instance, distractors in some cases looked similar to the targets, especially at longer distances (i.e., Humvee vs. another large vehicle). The effect of target-distractor similarity could have led to the need for increased visual attention to confidently distinguish between targets and distractors and decreased search efficiency (Duncan and Humphreys, 1989; Wolfe, 1994b, 2007; Zelinsky, 2008; Horstmann et al., 2019). It should also be noted that novelty of an object could have increased frequency of fixations. For instance, we would expect to see a difference in the Mean Number of Fixations and Mean Dwell Time for the Aircraft Condition and the Furniture Condition who had targets that varied in characteristics and models compared to the Humvee Condition and Motorcycle Condition with a target that stayed the same throughout the environment and only change in position and orientation in the environment. Subjects with a variable target may have fixated on more objects in general to determine if they should be included in their target count. Previous work has shown a disproportionate increase in visual attention on distractors for searches involving multiple targets compared single, static targets (Menneer et al., 2012). Novelty of the target can increase the time it takes to identify the object as a target among (varied) distractors (Lubow and Kaplan, 1997). The effect of target variation was not assessed in the current report due to low subject recruitment numbers in Target Conditions with a varied target. However, similar to previous work with multiple targets, we would expect that those with variable targets may have heightened attention toward distractors, negatively impacting their visual search efficiency throughout the task. Target characteristics, such as target variation and target-distractor similarity, may have been contributing factors to the large Number of Fixations reported overall.

Consideration of Contributing Factors Due to Task Design

Inherent differences in visual objects' shape, color, and size should have impacted visual attention toward specific objects in the virtual environment. However, rather than seeing these as limitations we argue that these are opportunities for additional, more nuanced research to better understand how: size, shape, color, visibility, context, etc. interplay with gaze behavior in ecologically valid environments. One would expect a greater Number of Fixations (and therefore greater Dwell Time) on the larger objects (e.g., Humvee, larger aircrafts, trucks, and buildings) compared to the smaller objects (e.g., furniture, motorcycles) due to being potentially visible at further distances. In contrast, a smaller object may be occluded by other larger objects or scenery until the subject is close to that object. In fact, we found that increased object surface size in the virtual

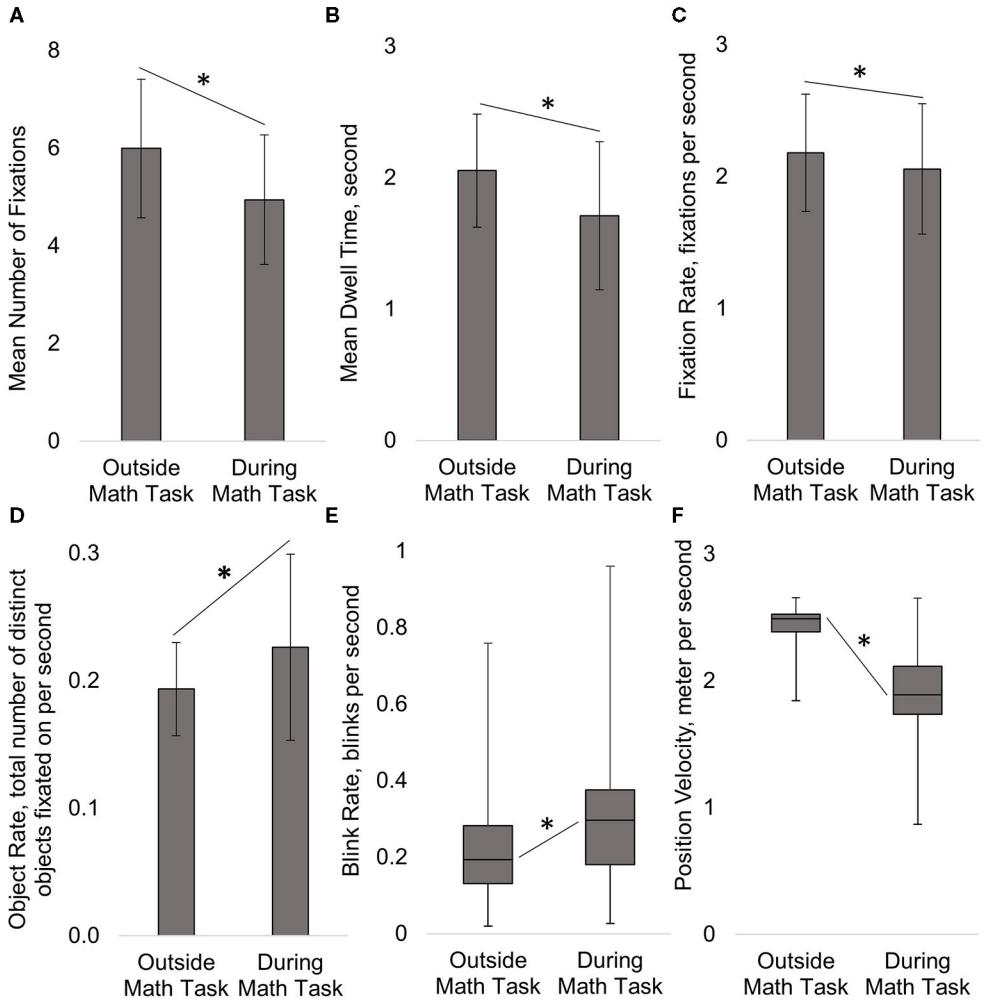


FIGURE 8 | During the Math Task subjects significantly decreased the Mean Number of Fixations **(A)**, Mean Dwell Time **(B)**, and Fixation Rate on objects **(C)**, but increased the number of unique objects fixated on per unit of time, Object Rate **(D)** and Blink Rate **(E)**. Subjects significantly decreased their velocity navigating the environment **(F)** during the Math Task. The duration of individual fixations and the proportion of fixations on objects as opposed to terrain or sky, did not significantly change between time periods (not shown in figure). Mean ± Standard Deviation (error bars) shown on figures **(A–D)** and Median ± Interquartile Range (error bars) on figure **(E,F)**. *$p<0.05$.

environment was significantly and positively correlated with the Mean Number of Fixations, the Mean Dwell Time, and the Mean Distance from the object when the fixation occurred (see Materials and Methods). This was also evident in our additional analysis looking specifically how the Humvee and Motorcycle Conditions looked at Humvees and motorcycles. Overall, Humvees had significantly greater Number of Fixations compared to motorcycles. Furthermore, it is interesting to note that those in the Motorcycle Condition devoted a greater Number of Fixations to this large distractor object compared to what the Humvee Condition devoted to the smaller distractor object. Although there were a greater Number of Fixations devoted to the Humvee target overall, it is also interesting to note that there was not a significant difference in overall Mean Dwell Time. It appears that the duration of these additional fixations was rather short and, perhaps, unintentional or the object was not of real visual interest. Therefore, it could be that subjects naturally fixated more on the larger objects, even if such objects were not the target assigned to them and not relevant to their assigned task (Võ and Wolfe, 2012). This may have also given those assigned to Target Conditions with the larger targets, the Aircraft Condition and the Humvee Condition, a distinct advantage in seeing their targets due to visibility.

Along with size, visibility in terms of where the object was physically placed in the environment, may have also driven visual attention toward or away from some objects. Objects were sporadically placed throughout the environment and items placed at the end of long stretches of the path may have been central to subjects' attentional locus while navigating down the path toward trail markers. These items, especially ones that

were centrally located along the horizontal plane, may have naturally drawn more visual attention (Karacan et al., 2010; Foulsham et al., 2011), especially if they were a larger object. For example, we found a surprisingly high number of fixations (∼16.5 fixations) and dwell time (∼5.9 s) on a particular GMC truck located at the end of a long canyon before a tight turn (compared to 7.1 fixations and 2.6 s averaged for all objects). When examined further, this particular object also had the highest Mean Distance (∼107 m) compared to the overall Mean Distance all objects in the environment (∼40 m). Therefore, some subjects could have fixated items due to their semi-random placement in the virtual environment rather than the due to the attributes of the item itself.

Scene context may also have impacted gaze toward certain objects in the virtual environment. For instance, the virtual environment was modeled as an arid and mountainous outdoor environment, but included some out of context items such as indoor furniture, musical instruments, a pool table, and a Ferris wheel. Scene context has shown to impact eye movement such as search time (Loftus and Mackworth, 1978; Henderson et al., 1999; Castelhano and Heaven, 2010) and memory recall (Draschkow et al., 2014). Items such as these may have garnered more visual attention due to their unexpected inclusion in the landscape (especially at the onset of the task) and/or could have been filtered as non-relevant visual objects if not assigned as a target that included those objects.

To help account for expected visual bias toward larger objects, random placement, or out of context objects in the virtual environment, we "normalized" each fixation metric for every object by subtracting the global mean for that object (the averaged value across all subjects for that particular object in the virtual environment). Normalization by simply dividing each gaze data point by the size of object (either 3D volume or 2D profile) in the virtual environment, resulted in a large bias toward the smaller targets. In contrast, our method of normalization enabled us to investigate object-centered gaze behavior for individuals compared to the mean across all conditions for any particular object. If in fact such bias was the cause of the increase in Mean Number of Fixations in the additional Humvee and Motorcycle Condition analysis for the Humvee object compared to the Motorcycle Condition, the normalization technique appeared to correct for such bias as differences were not present when using the Normalized Number of Fixations metric (**Figure 7** and **Table 3**).

Discrepancy With Virtual Environment and Real Life Walking Scenario in Distance of Focus

Mean distance in the virtual environment was around 40 m, with fixations on targets occurring at closer distances than distractors. As noted previously, our virtual environment allowed subjects to view objects down the path or to look around to their surroundings. Here, subjects appeared to fixate on objects relatively further away in their environment, which was previously noted for studies measuring gaze in a virtual setting (Clay et al., 2019). However, we would expect there to be some discrepancy between our findings and what occurs in real-world ambulation. Foulsham et al. (2011) found that people focus on objects further away in the view field when watching a first person video walking through an environment, compared to when they walked that environment in real life. In an ambulatory scenario, gaze is more often focused on near-field objects or terrain that could potentially affect gait. In a virtual environment navigation, gait perturbation is not a factor, thus near-field obstacles may be "under viewed" compared to what would occur in the real world.

Effect of a Divided Attention Task on Gaze Data

During the Math Task, there was a significant shift in subjects' eye movement behavior resulting from the increase in cognitive load. We found that subjects focused on more objects per second during Math Task, not by increasing Fixation Rate or shortening duration of each individual fixation, but by decreasing the Mean Number of Fixations on each object and therefore, total time spent on processing each object. Subjects also slowed down their navigation speed (∼24% decrease) and increased their Blink Rate (∼46%) during the Math Task. Additionally, the Proportion of Fixations on Objects in the virtual environment as opposed to those fixations on terrain or sky did not change significantly when the auditory task was present. Therefore, subjects did not appear to alter their visual attention away from objects and drift toward more background items in the environment (terrain and sky). Together these results suggests that subjects appeared to compensate for increased cognitive load by reducing the object processing time, slowing their physical pace of progression through the environment, and increasing their Blink Rate.

The change in subjects' eye movement behavior are consistent with previous work showing a tendency to give attentional preference to auditory stimuli, potentially at the cost of one's visual processing capabilities (Robinson and Sloutsky, 2010; Dunifon et al., 2016) and an increase in Blink Rate (Magliacano et al., 2020). Neurophysiological work with EEG has shown that when auditory stimuli are paired with a visual task (cross-modal processing) there is a latency in the visual P300 response but no negative impact on the processing of auditory stimuli (Robinson et al., 2010). Buetti and Lleras (2016) found that when subjects were asked to complete an auditory math task while looking at a screen passively, that subjects showed a decreased response to visual events (appearance of an image) on the screen, suggesting a decreased sensitivity to visual events. These findings are consistent with the decrease in object processing (decreased Number of Fixations and Dwell Time) found in the current study. One reason we may have seen a decrease in the Number of Fixations during the Math Task, was that the number of blinks increased. The increase in Blink Rate is consistent with findings from Magliacano et al. (2020) where they found an increase in Blink Rate accompanying an auditory counting task with the absence of any visual task. Increased Blink Rate has also been found to coincide with visual scenes that

require less attention and blinks are suppressed to reduce the chance of missing important information when visual attention is in demand (Nakano et al., 2009). Therefore, it is possible in our study that subjects disengaged from the visual task during the auditory math task, as evidence particularly by our significantly increased Blink Rate, due to the attention demand being comparatively low in the untimed visual search task. Due to the task design, we did not examine search efficiency in terms of a difference in fixations on targets and distractors during the Math Task, due to the auditory task occurring based on time in the environment (∼8 min mark) and not physical place in the environment where target and distractor appearance could be controlled for all subjects.

Visual attentional demands during the task due appear to be important in attentional compensation strategy when an auditory math task is simultaneously introduced. When combining an auditory divided attention task with a visual mismatch detection task (find the mismatch as soon as possible), Pomplun et al. (2001) found reduced efficiency in completing the visual task when the auditory task was also present, seen as increased task reaction time (detection of mismatch), Number of Fixations and Dwell Time. Thus, the visual compensation strategy adopted when the auditory stimuli is present, may depend on the degree of continuous response required for the visual task at hand. Our findings may contradict those found by Pomplun et al. (2001) study, perhaps due low visual attentional demands required during our task compared to a more timed and speeded-response task. Our active navigation (exploratory and self-paced) visual search task required the identification of targets from distractor objects and only required subjects to continually identify and keep a mental count, not provide a continuous response within a tight time constraint. Thus, in our study, subjects could shift task priority from performance in the visual search task to the Math Task without any immediate negative consequence. However, verbal responses from some subjects post-study did indicate they were challenged in remembering multiple mental summations simultaneously (summation of the Math Task problems and keeping the target count) indicating that the co-occurrence of the Math Task with the visual search task did have an impact on cognitive load overall.

It should also be noted that our findings differ from previous work where cognitive workload was increased by adding to the difficulty of the visual task itself (with no auditory input). Others have found that as a visual task becomes more complex and difficult, there is an increase in the Mean Number of Fixations (King, 2009; Buettner, 2013; Zagermann et al., 2018), an increase in Dwell Time (duration of fixations) (King, 2009; Meghanathan et al., 2015), an increase in the number of saccades (Zelinsky and Sheinberg, 1997; Zagermann et al., 2018), an increase in saccade rate (Buettner, 2013), and a decrease in Blink Rate (Benedetto et al., 2011; Maffei and Angrilli, 2018) during the completion of that visual task. Therefore, how cognitive load is increased in the study design is, once again, important to consider when examining the effects of increased cognitive load on eye movement metrics. Overall, our findings provide additional insight into the effect of an additional auditory task during a self-paced visual search task in a natural virtual environment.

Limitations

We would like to recognize several potential limitations to our study. One limitation was a restriction in data collection efforts due to public health concerns; we had to cease data collection earlier than planned and so were unable to have a balanced number of subjects in each Target Condition. This resulted in limited capabilities for comparison among the Target Conditions and their targets during the analysis. Second, while our experimental setup is similar to that of other studies, we utilized a desktop virtual environment instead of a virtual reality (VR) experience with a head mounted display. Although a VR system would provide a more immersive environment and allow for more free range in head and body movement compared to the current configuration, VR technology impose additional constraints when combining with other physiological measures, such as EEG. Likewise, simulator sickness is a common problem with immersive environments and our simulator sickness scores were relatively high overall. Simulator sickness could have impacted subject's natural viewing process through an environment and act as an unintended distractor from the task. Additionally, during the Math task it was observed that some subjects paused navigation when listening to the auditory number presentation (∼1–5 s), contrary to instruction and encouragement from experimenters. Therefore, gazed behavior during this time would be a reflection of cognitive processing and not necessarily the visual search and navigation task. Furthermore, there was no auditory simulation provided outside of that provided during the Math Task. Therefore, differences in eye movement could also be attributed to simple auditory processing and not necessarily due to increased cognitive load from the Math Task itself. Therefore, future work in with this study design should include a passive auditory stimulation throughout the navigation to truly examine cognitive load effects on this virtual search task in this virtual environment. Finally, in the current study we did not investigate any temporal patterns in gaze metrics, such as the change in number of re-fixations and Dwell Time on targets and distractors over time as subjects progressed through the virtual environment. Such temporal patterns have previously been investigated when investigating efficiency during hybrid target search in static scenes (Drew et al., 2017). It may also be interesting to see how gaze metrics change temporally as a function of physical distance from the object in the environment (i.e., the distribution of fixations with respect to distance from the object). Such future work would provide a more complete picture of how subjects' search efficiency changed over time and space in the environment.

CONCLUSION

In conclusion, we found that even during a self-paced navigation of a complex virtual environment, eye movement data can be used to robustly identify task-relevant gaze behaviors. There was a significant relationship between a subject's

gaze behavior (Number of Fixations and Dwell Time), their Target Condition, and objects in the environment. When an additional auditory Math Task was introduced, subjects slowed their speed, decreased the Number of Fixations and Dwell Time on objects in the environment, increased Blink Rate, and increased the number of objects scanned in the environment. The present study adds to our understanding of how individuals actively search for information while navigating a naturalistic environment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board, ARL 19-122. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LE was the primary individual on data processing, data analysis, and manuscript composition. SG also assisted in data processing. JT and SG contributed to the design and implementation of the original research. AR, JT, and SG contributed efforts and feedback on the analysis of the results and to the writing of the manuscript. RS wrote the software package for this study and provided feedback on the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was sponsored by the Army Research Laboratory and was accomplished under Contract Number W911NF-10-D-0002.

ACKNOWLEDGMENTS

We would like to thank Ashley Oiknine (AO), Bianca Dalangin (BD), and Min Wei (MW). Specifically, AO and BD for their work with subject recruitment and data collection, and MW for developing the initial world and task.

REFERENCES

Andersen, N. E., Dahmani, L., Konishi, K., and Bohbot, V. D. (2012). Eye tracking, strategies, and sex differences in virtual navigation. *Neurobiol. Learn. Mem.* 97, 81–89. doi: 10.1016/j.nlm.2011.09.007

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., and Montanari, R. (2011). Driver workload and eye blink duration. *Transport. Res. Part F Traffic Psychol. Behav.* 14, 199–208. doi: 10.1016/j.trf.2010.12.001

Buetti, S., and Lleras, A. (2016). Distractibility is a function of engagement, not task difficulty: Evidence from a new oculomotor capture paradigm. *J. Exp. Psychol. Gen.* 145, 1382–1405. doi: 10.1037/xge0000213

Buettner, R. (2013). "Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology," in *KI 2013: Advances in Artificial Intelligence, Lecture Notes in Computer Science,* eds I. J. Timm and M. Thimm (Berlin; Heidelberg: Springer), 37–48. doi: 10.1007/978-3-642-40942-4_4

Castelhano, M. S., and Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attent. Percept. Psychophys.* 72, 1283–1297. doi: 10.3758/APP.72.5.1283

Clay, V., König, P., and König, S. (2019). Eye tracking in virtual reality. *J. Eye Mov. Res.* 12, 1–18. doi: 10.16910/jemr.12.1.3

Davoudian, N., and Raynham, P. (2012). What do pedestrians look at at night? *Light. Res. Technol.* 44, 438–448. doi: 10.1177/14771535124 37157

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vis. Res.* 36, 1827–1837. doi: 10.1016/0042-6989(95)00294-4

Diaz, G., Cooper, J., Kit, D., and Hayhoe, M. (2013). Real-time recording and classification of eye movements in an immersive virtual environment. *J. Vis.* 13:5. doi: 10.1167/13.12.5

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., Kliegl, R., and Psychologie, A. (2011). Co-registration of eye movements and EEG in natural reading: analyses and review. *Analy. Rev. J. Exp. Psychol. Gen.* 140, 552–572. doi: 10.1037/a0023885

Draschkow, D., Wolfe, J. M., and Võ, M. L. H. (2014). Seek and you shall remember: scene semantics interact with visual search to build better memories. *J. Vis.* 14, 1–18. doi: 10.1167/14.8.10

Drew, T., Boettcher, S. E. P., and Wolfe, J. M. (2017). One visual search, many memory searches: an eye-tracking investigation of hybrid search. *J. Vis.* 17, 1–10. doi: 10.1167/17.11.5

Dukic, T., Ahlstrom, C., Patten, C., Kettwich, C., and Kircher, K. (2013). Effects of electronic billboards on driver distraction. *Traffic Inj. Prev.* 14, 469–476. doi: 10.1080/15389588.2012.731546

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96, 433–458. doi: 10.1037/0033-295X.96.3.433

Dunifon, C. M., Rivera, S., and Robinson, C. W. (2016). Auditory stimuli automatically grab attention: Evidence from eye tracking and attentional manipulations. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 1947–1958. doi: 10.1037/xhp0000276

Engbert, R., and Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vis. Res.* 43, 1035–1045. doi: 10.1016/S0042-6989(03)00084-1

Engbert, R., and Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7192–7197. doi: 10.1073/pnas.0509557103

Fotios, S., Uttley, J., and Yang, B. (2015). Using eye-tracking to identify pedestrians' critical visual tasks. Part 2. Fixation on pedestrians. *Lighting Res. Technol.* 47, 149–160. doi: 10.1177/1477153514522473

Foulsham, T., and Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world? *Can. J. Exp. Psychol.* 71, 172–181. doi: 10.1037/cep0000125

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vis. Res.* 51, 1920–1931. doi: 10.1016/j.visres.2011.07.002

Gidlöf, K., Wallin, A., Dewhurst, R., and Holmqvist, K. (2013). Using eye tracking to trace a cognitive process: gaze behaviour during decision making in a natural environment. *J. Eye Move. Res.* 6, 1–14. doi: 10.16910/jemr.6.1.3

Grüner, M., and Ansorge, U. (2017). Mobile eye tracking during real-world night driving: a selective review of findings and recommendations for future research. *J. Eye Mov. Res.* 10:1. doi: 10.16910/jemr.10.2.1

Helbing, J., Draschkow, D., and Vo, M. L.-H. (2020). Semantic and syntactic anchor object information interact to make visual search in immersive scenes efficient. *J. Vis.* 20:573. doi: 10.1167/jov.20.11.573

Henderson, J. M., Weeks, P. A., and Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 210–228. doi: 10.1037/0096-1523.25.1.210

Hoffman, J. E., and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Percept. Psychophys.* 57, 787–795. doi: 10.3758/BF03206794

Hollingworth, A., and Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 113–136. doi: 10.1037/0096-1523.28.1.113

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Weijer, J., et al. (2011). *Eye Tracking A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press. Available online at: http://lup.lub.lu.se/record/1852359 (accessed March 8, 2021).

Horstmann, G., Ernst, D., and Becker, S. (2019). Dwelling on distractors varying in target-distractor similarity. *Acta Psychol.* 198:102859. doi: 10.1016/j.actpsy.2019.05.011

Jangraw, D. C., Johri, A., Gribetz, M., and Sajda, P. (2014). NEDE: An open-source scripting suite for developing experiments in 3D virtual environments. *J. Neurosci. Methods* 235, 245–251. doi: 10.1016/j.jneumeth.2014.06.033

Kafkas, A., and Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *Q. J. Exp. Psychol.* 64, 1971–1989. doi: 10.1080/17470218.2011.588335

Karacan, H., Cagiltay, K., and Tekman, H. G. (2010). Change detection in desktop virtual environments: an eye-tracking study. *Comput. Hum. Behav.* 26, 1305–1313. doi: 10.1016/j.chb.2010.04.002

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220. doi: 10.1207/s15327108ijap0303_3

King, L. A. (2009). "Visual navigation patterns and cognitive load," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* Vol. 5638, eds D. D. Schmorrow, I. V. Estabrooke, and M. Grootjen (Berlin; Heidelberg: Springer), 254–259. doi: 10.1007/978-3-642-02812-0_30

Kingstone, A., Smilek, D., Ristic, J., Kelland Friesen, C., and Eastwood, J. D. (2003). Attention, researchers! It is time to take a look at the real world. *Curr. Direct. Psychol. Sci.* 12, 176–180. doi: 10.1111/1467-8721.01255

Kit, D., Katz, L., Sullivan, B., Snyder, K., and Ballard, D. (2014). Eye movements, visual search and scene memory, in an immersive virtual environment. *PLoS ONE* 9:94362. doi: 10.1371/journal.pone.0094362

Kothari, R., Yang, Z., Kanan, C., Bailey, R., Pelz, J. B., and Diaz, G. J. (2020). Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities. *Sci. Rep.* 10:2539. doi: 10.1038/s41598-020-59251-5

Kothe, C. (2014). *Lab Streaming Layer (LSL)*. Available online at: https://github.com/sccn/labstreaminglayer (accessed July 20, 2020).

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. (1995). The role of attention in the programming of saccades. *Vis. Res.* 35, 1897–1916. doi: 10.1016/0042-6989(94)00279-U

Land, M. F., and Lee, D. N. (1994). Where we look when we steer. *Nature* 369, 742–744. doi: 10.1038/369742a0

Lappi, O., Rinkkala, P., and Pekkanen, J. (2017). Systematic observation of an expert driver's gaze strategy-an on-road case study. *Front. Psychol.* 8:620. doi: 10.3389/fpsyg.2017.00620

Li, C. L., Aivar, M. P., Kit, D. M., Tong, M. H., and Hayhoe, M. M. (2016). Memory and visual search in naturalistic 2D and 3D environments. *J. Vis.* 16:9. doi: 10.1167/16.8.9

Liao, H., Dong, W., Huang, H., Gartner, G., and Liu, H. (2019). Inferring user tasks in pedestrian navigation from eye movement data in real-world environments. *Int. J. Geograph. Inform. Sci.* 33, 739–763. doi: 10.1080/13658816.2018.1482554

Livingstone-Lee, S. A., Murchison, S., Zeman, P. M., Gandhi, M., van Gerven, D., Stewart, L., et al. (2011). Simple gaze analysis and special design of a virtual Morris water maze provides a new method for differentiating egocentric and allocentric navigational strategy choice. *Behav. Brain Res.* 225, 117–125. doi: 10.1016/j.bbr.2011.07.005

Loftus, G. R., and Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 565–572. doi: 10.1037/0096-1523.4.4.565

Lubow, R. E., and Kaplan, O. (1997). Visual search as a function of type of prior experience with target and distractor. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 14–24. doi: 10.1037/0096-1523.23.1.14

Maffei, A., and Angrilli, A. (2018). Spontaneous eye blink rate: an index of dopaminergic component of sustained attention and fatigue. *Int. J. Psychophysiol.* 123, 58–63. doi: 10.1016/j.ijpsycho.2017.11.009

Magliacano, A., Fiorenza, S., Estraneo, A., and Trojano, L. (2020). Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neurosci. Lett.* 736:135293. doi: 10.1016/j.neulet.2020.135293

Matthis, J. S., Yates, J. L., and Hayhoe, M. M. (2018). Gaze and the control of foot placement when walking in natural terrain. *Curr. Biol.* 28, 1224–1233.e5. doi: 10.1016/j.cub.2018.03.008

Meghanathan, R. N., van Leeuwen, C., and Nikolaev, A. R. (2015). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Front. Hum. Neurosci.* 8:1063. doi: 10.3389/fnhum.2014.01063

Menneer, T., Stroud, M. J., Cave, K. R., Li, X., Godwin, H. J., Liversedge, S. P., et al. (2012). Search for two categories of target produces fewer fixations to target-color items. *J. Exp. Psychol. Appl.* 18, 404–418. doi: 10.1037/a0031032

Mueller, S. C., Jackson, C. P. T., and Skelton, R. W. (2008). Sex differences in a virtual water maze: an eye tracking and pupillometry study. *Behav. Brain Res.* 193, 209–215. doi: 10.1016/j.bbr.2008.05.017

Nakano, T., Yamamoto, Y., Kitajo, K., Takahashi, T., and Kitazawa, S. (2009). Synchronization of spontaneous eyeblinks while viewing video stories. *Proc. R. Soc. B Biol. Sci.* 276, 3635–3644. doi: 10.1098/rspb.2009.0828

Olk, B., Dinu, A., Zielinski, D. J., and Kopper, R. (2018). Measuring visual search and distraction in immersive virtual reality. *R. Soc. Open Sci.* 5, 1–15. doi: 10.1098/rsos.172331

Ouerhani, N., Von Wartburg, R., Hugli, H., and Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *ELCVIA Electron. Lett. Comput. Vision Image Anal.* 3, 13–24. doi: 10.5565/rev/elcvia.66

Palmer, J. (1995). Attention in visual search: distinguishing four causes of a set-size effect. *Curr. Dir. Psychol. Sci.* 4, 118–123. doi: 10.1111/1467-8721.ep10772534

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Front. Hum. Neurosci.* 9:660. doi: 10.3389/fnhum.2015.00660

Pomplun, M., Reingold, E. M., and Shen, J. (2001). Investigating the visual span in comparative search: the effects of task difficulty and divided attention. *Cognition* 81, B57–B67. doi: 10.1016/S0010-0277(01)00123-8

Robinson, C. W., Ahmar, N., Sloutsky, V., Robinson, C. W., and Sloutsky, V. M. (2010). Evidence for auditory dominance in a passive oddball task Publication Date Evidence for auditory dominance in a passive oddball task. *Proc. Ann. Meet. Cogn. Sci. Soc.* 32, 2644–2649.

Robinson, C. W., and Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisc. Rev. Cogn. Sci.* 1, 135–141. doi: 10.1002/wcs.12

Schrom-Feiertag, H., Schinko, C., Settgast, V., and Seer, S. (2017). Evaluation of guidance systems in public infrastructures using eye tracking in an immersive virtual environment. *Spat. Cogn. Comput.* 17, 163–183. doi: 10.1080/13875868.2016.1228654

Smith, T. J., and Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *J. Vis.* 13:16. doi: 10.1167/13.8.16

Tatler, B. W., and Tatler, S. L. (2013). The influence of instructions on object memory in a realworld setting. *J. Vis.* 13:5. doi: 10.1167/13.2.5

Tobii Pro (2018). *Eye Tracker Data Quality Test Report: Accuracy, Precision and Detected Gaze Under Optimal Conditions-Controlled Environment v. 1.0-en-US*. Available online at: www.tobiipro.com

Võ, M. L. H., and Wolfe, J. M. (2012). When does repeated search in scenes involve memory? Looking at versus looking for objects in scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 23–41. doi: 10.1037/a0024147

Võ, M. L. H., Zwickel, J., and Schneider, W. X. (2010). Has someone moved my plate? The immediate and persistent effects of object location changes on gaze allocation during natural scene viewing. *Attent. Percept. Psychophys.* 72, 1251–1255. doi: 10.3758/APP.72.5.1251

Watson, M. R., Voloh, B., Thomas, C., Hasan, A., and Womelsdorf, T. (2019). USE: an integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. *J. Neurosci. Methods* 326:108374. doi: 10.1016/j.jneumeth.2019.108374

Williams, C. C., and Castelhano, M. S. (2019). The changing landscape: high-level influences on eye movement guidance in scenes. *Vision* 3:33. doi: 10.3390/vision3030033

Wolfe, J. M. (1994a). Visual search in continuous, naturalistic stimuli. *Vis. Res.* 34, 1187–1195. doi: 10.1016/0042-6989(94)90300-X

Wolfe, J. M. (1994b). Guided search 2.0 A revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238. doi: 10.3758/BF03200774

Wolfe, J. M. (2007). "Guided search 4.0: current progress with a model of visual search," in *Integrated Models of Cognitive Systems*, ed W. D. Gray (New York, NY: Oxford University Press), 99–119.

Wolfe, J. M. (2012). Saved by a log: how do humans perform hybrid visual and memory search? *Psychol. Sci.* 23, 698–703. doi: 10.1177/0956797612443968

Zagermann, J., Pfeil, U., and Reiterer, H. (2018). "Studying eye movements as a basis for measuring cognitive load," in *Conference on Human Factors in Computing Systems - Proceedings* (Montreal, QC), 1–6. doi: 10.1145/3170427.3188628

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychol. Rev.* 115, 787–835. doi: 10.1037/a0013118

Zelinsky, G. J., and Sheinberg, D. L. (1997). Eye movements during parallel-serial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 244–262. doi: 10.1037/0096-1523.23.1.244

Author Disclaimer: The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Conflict of Interest: LE, RS, and SG were employed by the company DCS Corp.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fpsyg.2021.686183

Is Pupil Activity Associated With the Strength of Memory Signal for Words in a Continuous Recognition Memory Paradigm?

Jorge Oliveira[1], Marta Fernandes[2], Pedro J. Rosa[1] and Pedro Gamito[1]*

[1] Digital Human-Environment Interaction Lab, Lusófona University, Lisbon, Portugal, [2] Psychology Unit, Universidade da Madeira, Funchal, Portugal

Research on pupillometry provides an increasing evidence for associations between pupil activity and memory processing. The most consistent finding is related to an increase in pupil size for old items compared with novel items, suggesting that pupil activity is associated with the strength of memory signal. However, the time course of these changes is not completely known, specifically, when items are presented in a running recognition task maximizing interference by requiring the recognition of the most recent items from a sequence of old/new items. The sample comprised 42 healthy participants who performed a visual word recognition task under varying conditions of retention interval. Recognition responses were evaluated using behavioral variables for discrimination accuracy, reaction time, and confidence in recognition decisions. Pupil activity was recorded continuously during the entire experiment. The results suggest a decrease in recognition performance with increasing study-test retention interval. Pupil size decreased across retention intervals, while pupil old/new effects were found only for words recognized at the shortest retention interval. Pupillary responses consisted of a pronounced early pupil constriction at retrieval under longer study-test lags corresponding to weaker memory signals. However, the pupil size was also sensitive to the subjective feeling of familiarity as shown by pupil dilation to false alarms (new items judged as old). These results suggest that the pupil size is related not only to the strength of memory signal but also to subjective familiarity decisions in a continuous recognition memory paradigm.

Keywords: pupillary response, recognition memory, memory strength, eye tracking, pupillometry

OPEN ACCESS

Edited by:
Julien Epps,
University of New South Wales, Australia

Reviewed by:
Michael Francis Bunting,
University of Maryland, College Park, United States
Chin-An Josh Wang,
National Central University, Taiwan
Cezary Biele,
National Information Processing Institute, Poland

****Correspondence:***
Jorge Oliveira
jorge.oliveira@ulusofona.pt

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *26 March 2021*
Accepted: *25 October 2021*
Published: *23 November 2021*

Citation:
Oliveira J, Fernandes M, Rosa PJ and Gamito P (2021) Is Pupil Activity Associated With the Strength of Memory Signal for Words in a Continuous Recognition Memory Paradigm? Front. Psychol. 12:686183. doi: 10.3389/fpsyg.2021.686183

INTRODUCTION

Pupillometry has long been used in cognitive science as a measure of cognitive activity (Sirois and Brisson, 2014). This relationship was established in the 1960s, with evidence for associations between pupillary response and psychological processes such as arousal (Hess and Polt, 1960) and short-term memory (Kahneman et al., 1968). This interest has increased rapidly ever since, mainly not only due to its recording simplicity and non-intrusiveness compared with electrophysiological measurements but also due to the automaticity of pupillary response, which is associated with autonomous nervous system activity (Steinhauer et al., 2004) being controlled in the brain by the

superior colliculus (Wang and Munoz, 2015) and the locus coeruleus norepinephrine system (Joshi et al., 2016; Lewandowska et al., 2019).

The increasing interest on the relationship between pupil activity and memory processing is found in more recent debates (Brocher and Graf, 2017; Kafkas and Montaldi, 2017), which is revealed by a pupil dilation effect to familiar stimuli compared with unfamiliar stimuli.

In recognition memory designs, stimuli are encoded in a learning or study phase, being subsequently recognized in a test phase, where the (old) stimuli are intermingled with (new) interference stimuli. Studies using pupillary activity as an index for memory typically found an increase in pupil size for correctly recognized "old" stimuli relative to correct rejections of "new" stimuli in study-test procedures (Heaver and Hutton, 2011; Kafkas and Montaldi, 2012). This is known as the pupil old/new effect (Võ et al., 2008; for a review, van der Wel and van Steenbergen, 2018), which is considered as an outcome of the strength of memory signal associated with the retrieval of declarative memory (Papesh et al., 2012).

Otero et al. (2011) aimed at understanding the cognitive processes underlying pupil old/new effects in recognition memory by conducting various experiments manipulating the strength of memory signal for deep vs. shallow encoded items. The results revealed that the pupil old/new effect was more pronounced for remembered words (deeper encoding) compared with known words (shallow encoding). Brocher and Graf (2016) also demonstrated pupil old/new effects irrespective of lexicality, word valence, and frequency. More importantly, weakening the memory trace across these experiments, either by repeating legal vs. pseudowords or asking participants to make speeded responses, led to a reduction in pupil old/new effects, suggesting that conditions weakening memory signal would affect pupillary response.

Kucewicz et al. (2018) measured pupil size during encoding and recall of word lists. The lists consisted of 12-word items that were sequentially presented on a computer screen in the study phase. A distractor task was included between the study and test phases for interference. In the test phase, the participants were asked to verbally recall the word lists as fast as possible within 30 s. The authors studied the time course of pupillary response throughout the experimental task to examine the pupil dynamics for successfully recalled items compared with forgotten items. At the encoding phase, the results revealed an initial constriction followed by a pupil dilation, which increased as the word items were actively retained in memory. Moreover, an increase in pupillary response was found during word recall with the following decrease in pupil size as word items were being recalled, described as being related to the retrieval of information from memory.

Magliero (1983) and van Rijn et al. (2012) have conducted pupillometry studies manipulating the retention interval to evaluate the association with memory strength, where they found that longer retention levels increased task-evoked pupil responses. van Rijn et al. (2012) repeated the presentation of word lists with retrieval cues of paired associates in four repetitions of test trials to study the effects of repetition on the pupillary response. The results were intriguing, suggesting that repetition of word lists decreased pupillary response at retrieval. The differences between short and long retention intervals decreased with the repetition of word lists. The overall results suggest an association with retrieval effort given the effects of retention interval and repetition of word lists, supporting the hypothesis that the magnitude of pupil dilation is associated with memory strength for individual items, but in a reversed pattern than the one observed in pupil old/new effect studies.

To further explore the pupil old/new effects, Kafkas and Montaldi (2015) found that pupil activity distinguished between objective (i.e., veridical old/new status of the item) and subjective (i.e., subjective old/new decision) familiarity and novelty in two distinct temporal components. One early component was found for the objective status, while a late component near the recognition response was found for the subjective status of items, which indicates that pupil activity may be sensitive to both explicit and implicit components of recognition memory.

This study evaluates the relationship between pupil activity and recognition memory in a running recognition task (Shepard and Teghtsoonian, 1961) with varying retention intervals to assess pupil activity during explicit manipulations of memory strength. In such a task, participants should retain information that is presented in a continuous sequence of items until the test trial for memory retrieval. This task may provide a more ecological way to assess human memory processing while maximizing interference compared with recognition memory of word lists where the study-test phases are separated by isolated interference tasks. This paradigm was used earlier in behavioral studies to manipulate the retention interval in visual word recognition (e.g., Shepard and Teghtsoonian, 1961; Coney and MacDonald, 1988; Federmeier and Benjamin, 2005), but this is the first study to use the continuous recognition memory paradigm in pupil research. According to the strength account, we would expect the recognition performance and pupil dilation to decrease as the retention interval increases. Our intent is also to explore the pupil dynamics in a continuous recognition memory design by assessing pupillary responses to the objective and subjective old/new status of word items.

MATERIALS AND METHODS

Participants

The sample comprised 42 adult Portuguese native speakers who had normal vision or corrected-to-normal vision, mostly women (n = 23) with a mean age of 26 years (SD = 6.79) and no less than 12 years of formal education. The participants were selected in a university campus for voluntary participation in a study related to "visual perception and memory." The exclusion criterion was history of psychiatric disorder or medication/drug use. The initial pool comprised 47 participants, but five participants were excluded due to low quality (more than 50% of data loss) of pupillary recordings or due to problems in the collection of behavioral responses.

Materials and Design

The stimulus words were collected from a database of validated Portuguese words from a sample of undergraduate students (Marques et al., 2007). For this study, we selected 107 words of 4 to 7 letters in length: 64 of these were used as study words and 43 as "new" test words. Both lists of words were matched for psycholinguistic variables of familiarity and age of acquisition.

Design and Procedure

This study was approved by the ethics committee of the host institution where it was carried out. The experiment was conducted in a soundproof booth with a constant low-bright room during only one session. The visual word recognition task was based on a continuous recognition memory paradigm originally from Shepard and Teghtsoonian (1961), with study words presented two times in a study-test procedure. In our task, study words were repeated in the test phase, intermingled with (new) interference words with different retention intervals. All participants were tested with words presented at four different interval levels manipulated through the number of words between study and test: lag 1 (immediate repetition), lag 4 (4 words separating study-test phases), lag 8 (8 words), and lag 32 (longest lag with 32 words between the study-test phases).

Each trial in the study phase began with a fixation cross for 250 ms preceding the word stimulus that was on the screen for 1,750 ms. In the test phase, each trial began with a mask consisting of a row of seven symbols ("&&&&&&&") for 250 ms, being replaced by the word stimulus (1,750 ms), according to the design of Heaver and Hutton (2011). All stimuli were presented at the center of the screen. The word stimulus in the test phase was followed by the mask that remained on the screen until a response was given. The recognition responses were given at this stage. The participants were instructed to respond with the keypress only when the word stimulus was replaced by the mask and during the time, the mask was visible on the screen. Following each word in the test phase, the participants also had to indicate their level of confidence in the decision (1, not at all confident to 5, very confident). Each trial of the study phase consisted of the mask and the word stimulus, whereas in the test phase, word stimuli were replaced by the mask (where recognition response was given) followed by the confidence level screen. The interstimulus interval was 1,000 ms for both the study and test phases. This procedure was the same between the different retention intervals. The only difference between retention conditions was the number of intervening items between the study and test phases. Intervening items were the number of words in a continuous sequence that comprised study words and "old" and "new" test or interference words. An example of the continuous recognition memory procedure is shown in the following sequence, where each letter describes a different word and the question mark the test phase:

a b c a? c?

In this sequence, "a" is tested at a lag of 3 and "c" is tested at a lag of 2 words between the study and test phases. This design is also illustrated in **Figure 1**.

The words were presented in black capital letters (38-point Arial font) over a gray background screen (Red = 128, Blue = 128, and Green = 128).

After informed consent, each participant was seated at a distance of 60 cm from the infrared eye-tracking system (Tobii T60, Tobii Technology AB, Danderyd, Stockholm, Sweden; instrument noise, 0.06 RMS). The calibration of the eye tracker was carried out for each participant using a five-point calibration setup.

Participants were instructed to keep still to minimize data loss due to head and body movements during the task. Following this stage, the participants completed a 5-min preliminary practice stage using proper nouns as stimuli before the recognition task. They were instructed to indicate in the keyboard whether a word was old (previously seen during the experiment) or not, as fast as possible.

The visual word recognition task was designed in Superlab (version 1.0.2; Cedrus Corporation, San Pedro, CA, United States) and presented through the 17-inch monitor of the eye tracker with a 1,280 × 1,024 resolution. The behavioral measures were collected using Superlab, and pupil responses were registered in Tobii Studio (version 3.0; Tobii Technology AB, Sweden), which is the native application of Tobii eye trackers. Eye data of both eyes were collected at a sampling rate of 60 Hz.

Data Pre-processing

Raw pupil data were exported from Tobii Studio version 3.3.2 software to SPSS (Version 25.0. Armonk, NY: IBM Corp.) for data reduction. The proportion of the missing values was first analyzed to assess the noise in pupil data (missing data = 3.97%). Missing pupil data were randomly distributed across trials. Pupil amplitude artifacts (<1 or >9 mm), as well as drifts and blinks, were coded as missing values (Rosa et al., 2015). Pupil diameters of zero lasting between 100 and 600 ms were considered blinks (Cosme et al., 2021), and replaced using linear interpolation (Carvalho and Rosa, 2020). Finally, a seven-point weighted average filter was applied to smooth data. The data file was then exported to Vision Analyzer software (version 2.1; Brain Products GmbH, Germany) for data segmentation and estimation of evoked pupil responses. The epochs were created for each stimulus category with stimulus-locked segments of 4,000 ms in length (i.e., from −250 to 3,750 ms at stimulus onset). This segmentation resulted in 64 segments, 16 segments for the study words tested at each of the four retention levels, plus 43 segments for the "new" test words (interference words were presented only during the test phase), in a total of 107 segments. The words at retrieval were visible during the first 1,750 ms of this time window. The remaining interval between 1,750 and 3,750 ms comprised the recognition response.

Pupil responses were calculated within each time bin of 250 ms for a time window of 3,750 ms. The baseline was set at −250 ms before the stimulus onset. The percentage of variation relative to baseline was calculated to depict the amplitude of pupillary responses to each experimental condition.

The behavioral measures consisted of accuracy from the signal detection theory (SDT), which comprises hits, correct rejections, false alarms, and misses. According to the SDT, hits

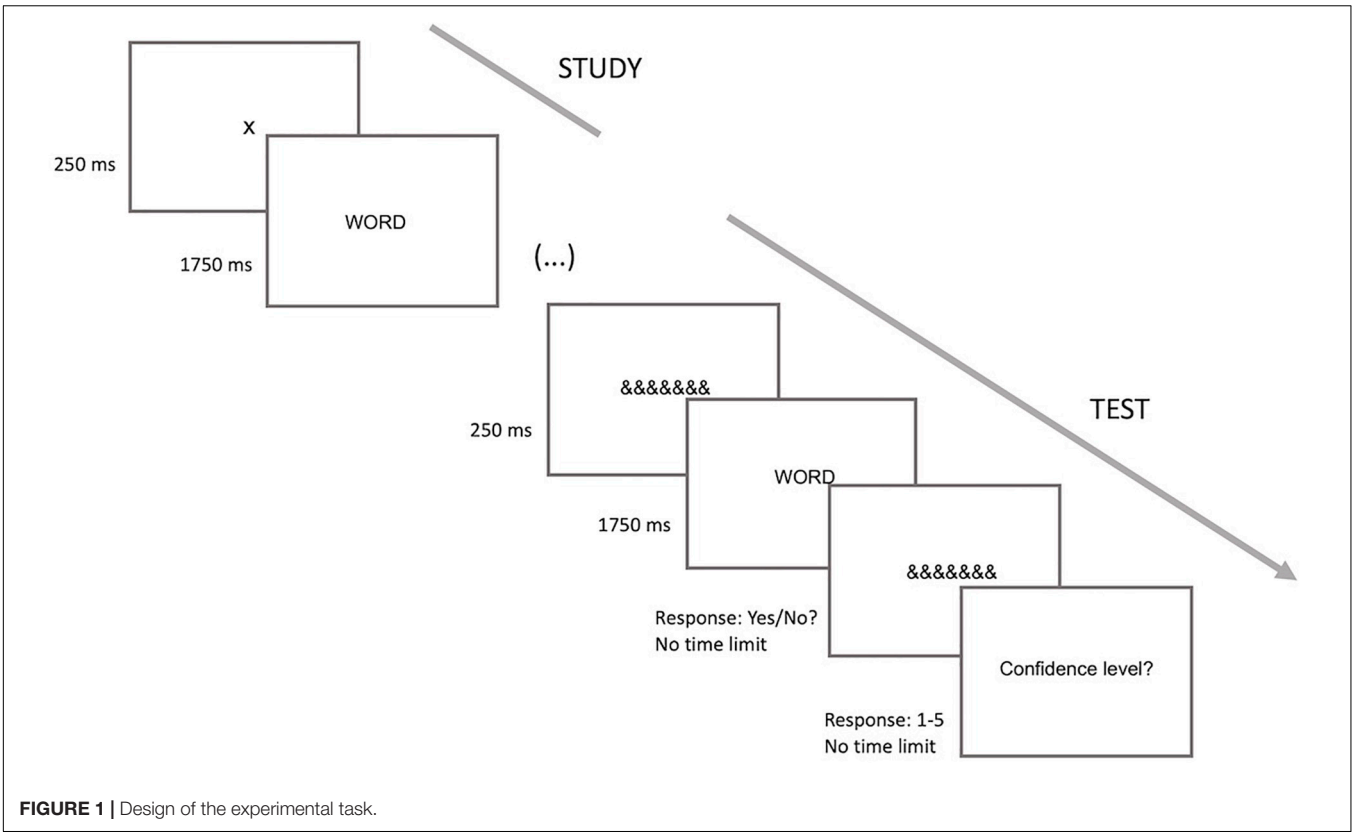


FIGURE 1 | Design of the experimental task.

and correct rejections depict correct decisions, whereas false alarms and misses are incorrect decisions that may be due to internal/external factors affecting human perception. Reaction times and confidence ratings were also assessed during this task.

RESULTS

Behavioral Measures

The analysis on behavioral measures was conducted for discrimination ability, reaction times, and confidence ratings in recognition responses. These variables were analyzed by retention intervals using repeated-measures ANOVA. Confidence levels were also assessed with receiver-operating characteristics (ROC) for determining the ability to distinguish recognition responses.

Recognition Accuracy

Recognition accuracy was calculated according to the SDT through *d*-prime (*d*') in which higher values describe better memory performance, which is given by the following expression: *d*' Z(H)-Z(FA). Participants had an average hit rate (correct recognition) of 81% (ranging from 37 to 100%) and a false alarm rate of 11% (ranging from 0 to 29%).

The effect of retention interval on recognition accuracy was analyzed with a single-factor repeated measures ANOVA with four levels (retention level: 1, 4, 8, and 32 items). The ANOVA showed significant differences with Greenhouse-Geisser correction in recognition accuracy between retention levels [F (1.430, 58.611) = 16.947; $p < 0.001$; $\eta^2_p = 0.292$], suggesting a significant decrease (Bonferroni corrected pairwise comparisons) from lag 1 to lag 4 ($p = 0.020$) and from lag 4 to lag 8 ($p = 0.002$). **Table 1** describes recognition performance in the running recognition task through *d*-prime, hits, false alarms, confidence levels, and reaction times across lag conditions. **Table 2** depicts the inference statistics for these analyses. The same pattern of results was observed for hits [F (2.551, 107.159) = 10.591; $p < 0.001$; $\eta^2_p = 0.201$] and false alarms [F (2.740, 115.86) = 3.698; $p < 0.05$; $\eta^2_p = 0.081$].

Confidence Ratings

The confidence levels in each of the recognition decisions were rated on a five-point Likert scale. The same design was used for the ANOVA that showed a similar pattern to that of the *d*-prime. These results indicated a decrease in confidence level for longer retention levels [F (2.168, 88.905) = 27.006; $p < 0.001$; $\eta^2_p = 0.397$]. Pairwise comparisons with Bonferroni correction indicated that the confidence level was highest in lag 1 and lowest in lag 32. Confidence level decreased from lag 1 to lag 4 ($p = 0.001$) and from lag 8 to lag 32 ($p < 0.001$).

A descriptive analysis on confidence ratings showed that most responses were extreme-confident responses. This pattern has limited further analyses between pupil data and confidence ratings, given the lack of valid cases in each cell for factorial designs. We have conducted a ROC analysis on confidence ratings to understand whether confidence would

TABLE 1 | Descriptive statistics for behavioral measures.

	Lag 1		Lag 4		Lag 8		Lag 32	
	M	***SE***	***M***	***SE***	***M***	***SE***	***M***	***SE***
d'	3.19	0.19	2.35	0.20	1.98	0.18	1.97	0.16
Hits	0.87	0.03	0.82	0.03	0.81	0.03	0.72	0.03
FA	0.14	0.02	0.15	0.03	0.20	0.03	0.19	0.02
CR	4.78	0.48	4.56	0.67	4.56	0.62	4.21	0.09
RT	1305.84	130.78	1613.07	142.80	1619.71	139.15	1612.46	147.53

M, mean; SE, standard error for the mean; d', d-prime for accuracy; CR, confidence ratings; RT, reaction times; and FA, false alarms.

TABLE 2 | Inference statistics for behavioral measures.

	MSE	η^2_p	***F***[a]	**Pairwise**[b]
d'	29.007	0.292	16.947***	l1 > l4 > l8,l32
Hits	0.190	0.201	10.591***	l1,l4 > l8,l32
FA	0.039	0.081	3.698*	l1,l4 < l8,l32
CR	3.190	0.397	27.006***	l1 > l4,l8 > l32
RT	1473240.24	0.143	6.836**	l1 < l4,l8,l32

MSE, mean square error; η^2_p, effect size through partial eta squared; and F, analysis of variance statistic; l1, lag 1; l4, lag 4; l8, lag 8; l32, lag 32.
[a]Greenhouse-Geisser correction. [b]Bonferroni corrected pairwise comparisons.
**p < 0.05; **p < 0.01; ***p < 0.001.*

discriminate successful recognition. This analysis showed a poor discriminant ability of confidence ratings on the recognition ability (AUC = 0.546; *SE* = 0.040; $p = 0.249$).

Reaction Time

Reaction time was also assessed through the same ANOVA to test the significant differences between lag conditions. The ANOVA revealed a significant difference in reaction times across retention levels [F (2.056, 84.298) = 6.836; $p = 0.002$; $\eta^2_p = 0.143$], with faster responses for words tested immediately at lag 1 that differed from the remaining conditions (all p's < 0.05).

Pupillometry

Pupil size analysis was performed in different steps. First, the analysis was conducted for pupillary responses to each lag condition. Second, the pupil old/new effect was calculated by comparing correct recognition responses to "old" words with correct rejections of "new" test words. Following these analyses, the pupillary responses were analyzed for recognition errors, namely, false alarms, i.e., incorrect rejections of new test words and misses, i.e., omissions in recognizing old words. The factor related to confidence levels in recognition was not included in the factorial design due to the insufficient number of trials for low confidence conditions, but this factor was controlled in further analyses by dividing the five-point Likert scale in a dichotomous variable for low and high confident decisions. Therefore, pupillary responses to false alarms were analyzed by confidence (low vs. high) to study whether the pupil activity is also associated with subjective familiarity (i.e., evaluating "new" test items as "old"). Finally, the pupillary responses across lag conditions were also studied for extreme-confident decisions (i.e., confidence rating equal to 5).

Pupil Dynamics by Retention Interval

Evoked pupillary responses for correct recognition decisions were analyzed to each retention condition (study-test lag) by plotting peak activity at 250 ms bins of the 3,750 ms time windows with a two-factor ANOVA. The retention level (4 levels) and bin (16 levels) were entered in this analysis as factors within-subjects.

The ANOVA revealed significant main effects for lag [F (1.718, 189.211) = 33.896; $p < 0.001$; $\eta^2_p = 0.453$] and bin [F (2.776, 189.211) = 23.939; $p < 0.001$; $\eta^2_p = 0.369$]. The main effect of lag described a decrease in pupil dilation for longer retention spans, whereas the main effect of bin described a pupil constriction at the initial stage of memory retrieval followed by a later dilation. This analysis also showed a significant interaction effect between factors [F (4.605, 189.211) = 5.949; $p < 0.001$; $\eta^2_p = 0.127$], suggesting a different pattern of pupil dynamics according to the retention condition. Pairwise comparisons (Bonferroni corrected) for retention level suggested a stronger pupil constriction for lags (all p's < 0.05) other than lag 1, and a later dilation for all retention conditions (all p's < 0.05). The differences were found mostly between lag 1 and the remaining lag conditions. This pattern is illustrated in **Figure 2**.

Pupil Old/New Effect

To further explore these results, the differences between evoked pupillary responses to "old" test words and "new" test words were calculated for studying the pupil old/new effect observed previously in recognition memory studies. The pupillary responses to each retention condition were compared with interference test words through a separate repeated measures ANOVA. The ANOVAs revealed the pupil old/new effect only at lag 1 [F (2.948, 120.883) = 6.972; $p < 0.001$; $\eta^2_p = 0.145$]. The results were also significant for the remaining retention levels but revealing a pupil constriction to "old" words compared with "new" words for lag 4 [F (3.123, 128.040) = 3.035; $p = 0.030$; $\eta^2_p = 0.069$], lag 8 [F (2.286, 93.725) = 4.169; $p = 0.014$; $\eta^2_p = 0.092$], and lag 32 [F (2.964, 121.504) = 3.343; $p = 0.022$; $\eta^2_p = 0.075$], as depicted in **Figure 3**.

Pupil Dynamics to Recognition Errors

Pupil activity was also analyzed for recognition errors. According to the SDT, the failure in detecting an item presented previously at the learning phase is defined as a miss, whereas the failure to reject a new item (interference word) is defined as a false alarm. The comparison with the repeated measures two-factor

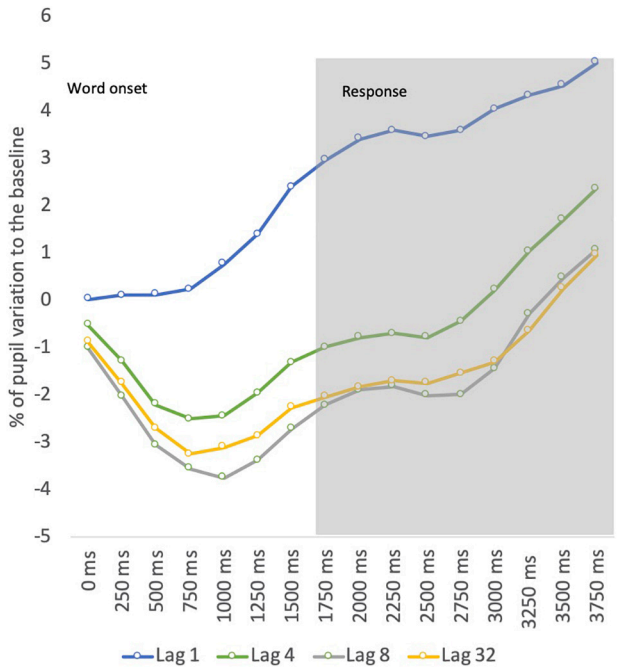


FIGURE 2 | Pupil dynamics by retention interval.

(type of recognition error and bin) ANOVA revealed a significant main effect, suggesting an overall difference in pupil dilation between misses and false alarms, with increased pupil dilation for false alarms [F (1, 77.123) = 6.806; p = 0.023; η^2_p = 0.233]. No interaction effects were found indicating that the pattern of pupil activity is not different between the two types of recognition errors (**Figure 4**).

Pupil Dynamics for False Alarms in High vs. Low Confident Decisions

Given the increased response to false alarms, in which the mean percentage of pupil dilation to the baseline was 2.10%, being very similar to the mean dilation observed for words tested at lag 1 (2.49%), we have conducted a further analysis by confidence levels (low vs. high) for false alarms to analyze pupil activity in subjective familiarity decisions. The comparisons between high-confident responses (confidence rating of 5) and less-confident responses (confidence rating below 5) in false alarms show a marginally significant difference [F (1, 34.965) = 4.663; p = 0.054; η^2_p = 0.298] between the mean dilation to high-confident responses (2.9%) and less-confident responses (−0.78%), as depicted in **Figure 5**.

Pupil Dynamics by Retention Interval for High-Confident Decisions

The above results suggest that pupil activity may be sensitive to subjective familiarity, which may occur when the participant rejects a "new" interference item probably being influenced by the subjective feeling of knowing that such an item was old. This may have been the case for extreme-confident decisions in false alarms. Therefore, the pupillary response by retention condition was reanalyzed only for extreme-confident decisions. The same two-factor ANOVA detected a main effect of retention condition [F (2.534, 101.376) = 20.328; $p < 0.001$; η^2_p = 0.337], showing the same temporal pattern across lag conditions. A main effect of bin was observed [F (3.068, 122.705) = 27.740; $p < 0.001$; η^2_p = 0.410], while the interaction effect [F (6.896, 275.846) = 4.868; $p < 0.001$; η^2_p = 0.108] revealed a decrease (Bonferroni corrected) in evoked pupillary responses across lag conditions providing similar results to that of the ANOVA without controlling for confidence ratings (**Figure 6**).

Pupillary Response by Retention Interval According to the Number of Interference Items

In this experimental task, the intervening items separating study-test trials comprised both study words, "old" studied words and "new" test (interference) words. Interference varied according to the number of "new" test words separating the study-test trials. This variable related to interference was divided according to the median for trials with low interference vs. high interference. This analysis was conducted with a two-factor repeated-measures ANOVA (retention level with 3 levels: 4, 8, and 32 items and interference: low vs. high). The retention level 1 was not included as this condition consisted of immediate recognition. The ANOVA did not reveal significant effects of interference in pupil dilation (all p's > 0.05), although the visual inspection to **Figure 7** suggests an interaction between interference and lag condition on pupil activity.

DISCUSSION

This study aimed to investigate the relationship between pupil activity and recognition memory according to explicit manipulations of memory strength in a continuous recognition

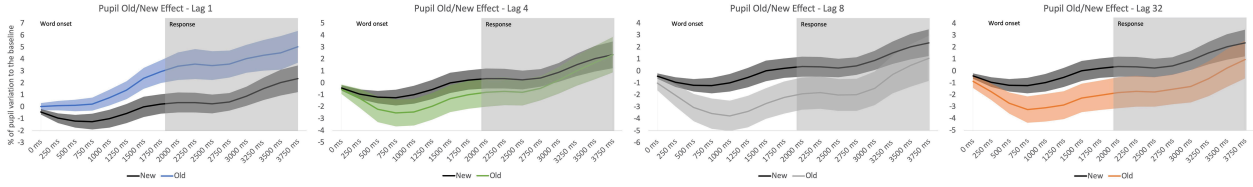


FIGURE 3 | Pupil old/new effect.

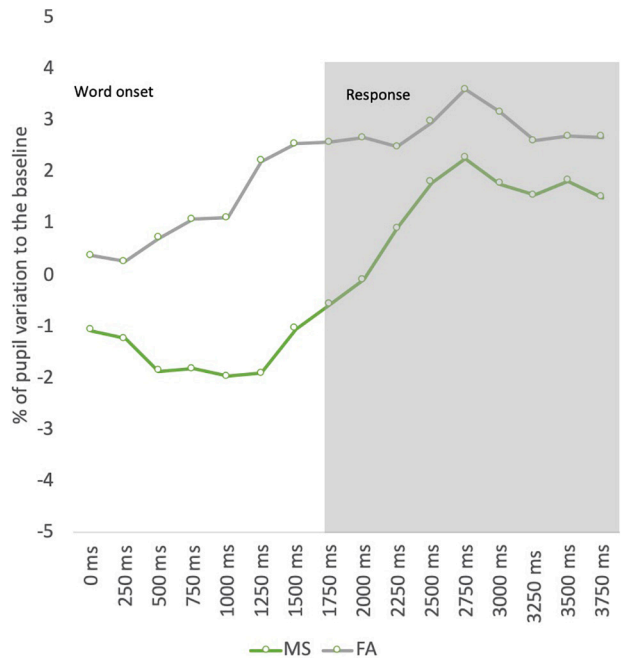


FIGURE 4 | Pupil dynamics to recognition errors.

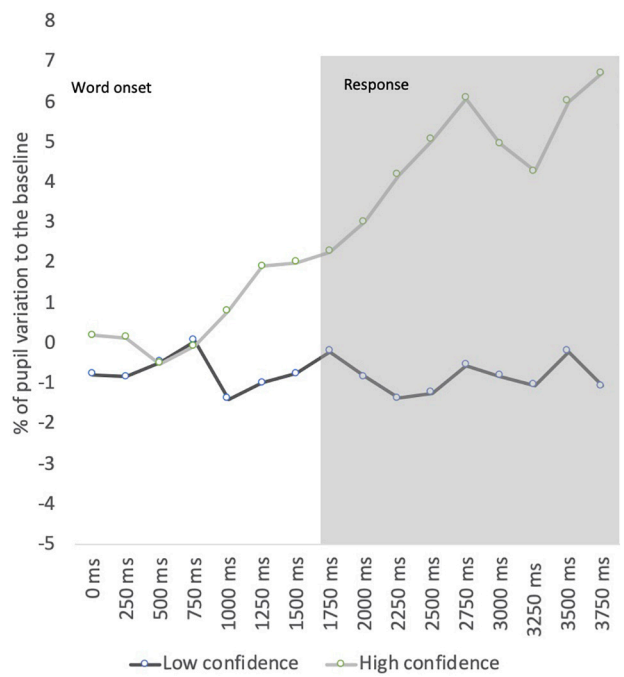


FIGURE 5 | Pupil dynamics for false alarms in high vs. low confident decisions.

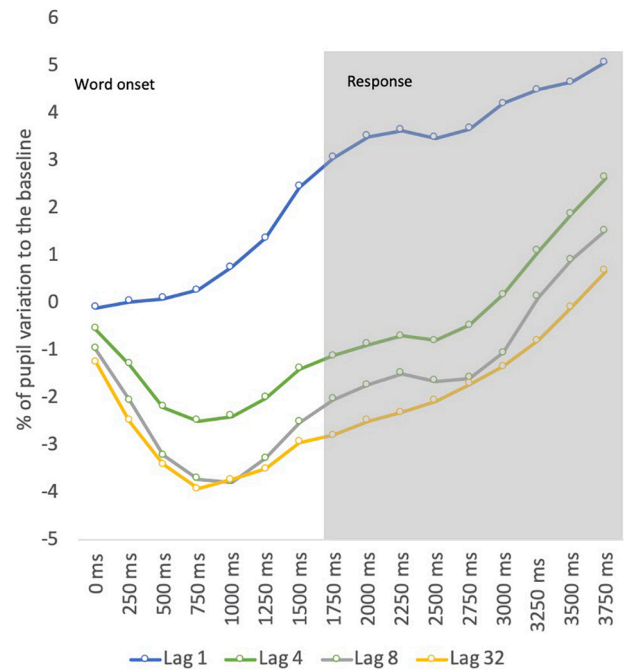


FIGURE 6 | Pupil dynamics by retention interval for high confident decisions.

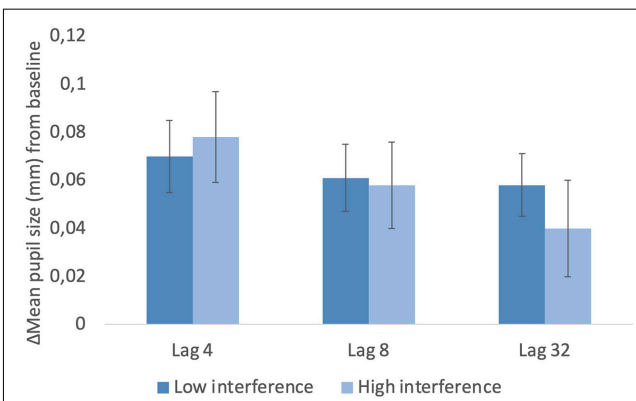


FIGURE 7 | Pupillary response by retention interval according to the number of interference items.

memory design. This goal was achieved by exploring pupil dynamics across different retention intervals to objective and subjective old/new status of word items in a running recognition task.

The behavioral data show a decrease in recognition performance with increasing retention intervals in recognition. The discrimination ability decreased with an increasing lag between study and test items, mostly in the transition from shorter retention (lag 1) to moderate retention (lag 4 and lag 8). Confidence ratings also decreased at longer retention levels, which distinguished shorter (lag 1), moderate (lag 4 and lag 8), and longer (lag 32) retention intervals. The results from reaction time were in the same direction but indicated an earlier impact in recognition performance from lag 1 to lag 4. Altogether, the behavioral results suggest that the recognition task was effective in manipulating memory strength as recognition performance decreased with an increasing lag between study and test of word

items, but the increase in reaction times also indicates that effort may have increased across the retention intervals.

The pupil data revealed increased pupil dilation for words tested at lag 1. Likewise, the pupil old/new effect was found only for words tested at the shortest retention interval. The comparison between lag 1 and the remaining lag conditions showed that mean pupil dilation decreased as retention levels increased. These data contradict previous studies on working memory that suggest an increase in pupillary response when the number of items maintained into memory increased up to 4–5 items (Unsworth and Robison, 2018). Therefore, if the current results depicted working memory processing, we should expect an increase in pupillary response at least until lag 4 (i.e., four words between study and test), instead of the decrease observed from lag 1 to lag 4.

Our data also revealed that differences in the pattern of pupil activity across lag conditions were evident mainly by stronger pupil constrictions to items recognized at longer retention intervals. Considering that each study trial lasts approximately 2 s, the retention interval between the study and test phases for a stimulus tested at lag 4 is about 8 s, at lag 8 is about 16 s, and at lag 32 is about 64 s. Pupillary responses at lag 1 may correspond to a condition when the stimulus is still active in memory endorsing larger pupil dilations, comparing with longer retention levels when other memory processes may occur as an active rehearsal for long-term memory storage. This early constriction is not likely to be related to light reflex during the baseline period because in our study pupil baseline was calculated at 250 ms before the stimulus onset corresponding to a string of symbols to minimize the influence of luminance during the transition to the target stimulus while also preventing the accommodation effects on pupil size.

The study by van Rijn et al. (2012) also revealed an initial pupil constriction during word retrieval, but in our study, the size of this initial constriction seems to be associated with memory strength as this was more pronounced for items that were recognized at longer retention levels. In a previous study, using temporal analysis for pupillary response to complex stimuli (i.e., scenes) revealed that the initial constriction of pupil size during memory retrieval was related to novelty, where novel scenes elicited stronger pupil constrictions compared with familiar scenes in high confident decisions (Naber et al., 2013). In this study, this prediction was not possible to investigate as this would require novel items that were not familiar to the participants. In our study, we selected only high familiarity words to control for familiarity effects. A *post hoc* analysis to familiarity by splitting the data according to the median level of familiarity did not reveal significant effects on pupil data, although this result should be interpreted with caution given the low range of familiarity levels for item words used in this study, which varied from 1.1 to 3.5 for 4–7 letter words (Marques et al., 2007).

The decrease in pupil dilation across lag conditions contradicts the effort accounting that memory effort increases pupil dilation (e.g., Granholm and Steinhauer, 2004; van Rijn et al., 2012), as the increase in effort revealed by an increase in reaction times should have produced increased pupil dilations, but the reverse was found in our study. Another study found increased pupil dilation for study lists repeated once corresponding to a more effortful condition compared with items retrieved after more repetitions (van Rijn et al., 2012). One possible explanation for these differences may be related to the nature of the task employed in our study. In this running recognition task, performance at each retention interval may be affected not only by decay (time) but also by interference in an overall effect, which differs from tasks employing single lists of items that study words in isolation. The decrease found in pupil dilation across retention levels may be related to decay and interference as longer retention intervals imply more intervening items and longer periods of time between the study and test phases. The intervening items were words in a continuous sequence that comprised both study words, "old" studied words and "new" test or interference words, being the latter used to fill the sequence at each retention condition. To investigate whether interference through the number of interference words influenced pupil dilation, the test trials for each of the retention conditions were divided by the median number of interference items, which did not show significant effects on pupillary response. It is advisable that the future studies have to distinguish between the effects of decay (time) and the number of interference words in the recognition task. Moreover, the manipulation of repetition of test trials in an adapted version of this continuous recognition memory design will be crucial to study in more detail the effects of memory effort across lag conditions. The assessment of vigilance and fatigue levels will be also an important consideration for further studies. Despite this, recognition design may minimize the potential effects of fatigue, as the retention interval was randomly manipulated across the continuous recognition procedure, future studies should consider both *online* measures as eye blink analysis and *offline* self-reports for assessing fatigue levels in continuous recognition memory designs to better describe pupil activity.

Furthermore, the results were also explored regarding recognition errors. The data revealed that false alarms (new items judged as old) elicited an increased pupil dilation compared with misses (old items judged as new). These data are aligned with the results from Kafkas and Montaldi (2015) that found increased pupil dilations for false alarms compared with misses, which discriminated between an early component of pupil data reflecting the objective veridical status of old/new items and a late component reflecting the subjective status of old/new items. To explore whether the subjective recognition decision modulates pupillary response, our data were analyzed according to the confidence level in false alarms. The results indicate that pupils dilated more when participants believed a new item was previously seen during the sequence mainly for high-confident incorrect decisions. Nevertheless, the analysis of confidence effects in pupil size across the retention interval did not seem to modulate pupil response for correct decisions. This latter analysis may have been affected by the lack of sensitivity as most correct responses were accompanied by extreme-confident decisions. In fact, the ROC analysis shows that this variable did not discriminate recognition responses. Future studies should also use feasible confidence scales to distinguish confidence in recognition decisions more effectively.

In sum, these results point to a relationship between pupillary response with the strength of the underlying memory signal in light of the following data: (1) The increase in retention interval decreased overall pupil dilation; (2) the pupil old/new effect was evident only for the shortest retention level; and (3) the analysis on the dynamics of pupillary response revealed a different pattern of pupil activity across the retention interval. However, it is also important to note that this response may be dependent on the subjective feeling of familiarity to a given item, as pupil size was also modulated by incorrect recognition decisions to "new" interference words especially those with high confidence.

Given the simplicity and non-intrusiveness of a pupil size measurement, the development of reliable methods for assessing pupil activity may provide an ecologically valid measure for assessing human memory and behavior in complex environments. The integration of pupil size measurement in virtual reality environments need not wait for further research. For instance, Juvrud et al. (2018) have demonstrated that it is possible to have a method based on a virtual reality scenario for assessing pupillary responses not depending on low-level stimulus features. In such virtual reality environments, it will be interesting to explore the current results under naturalistic contexts using stimuli other than words (i.e., objects, faces) and test whether pupillary responses are associated with the strength of memory in conditions that resemble real-life situations. Likewise, the study of false memory in virtual reality environments will be also intriguing given the current results suggesting the sensitivity of pupillary response not only to the objective oldness of the items but also to the subjective feeling of familiarity that drives recognition decisions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comissão de Ética e Deontologia para a Investigação Científica – CEDIC. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JO was responsible for designing the study and writing the initial version of the manuscript. MF conducted data collection and contributed to the initial version of the manuscript. PR was responsible for data processing procedures, whereas PG was involved in the statistical analyses. All authors have contributed and approved the final version of the manuscript.

FUNDING

This study was supported by HEI-Lab, the research unit from Lusófona University, which is funded by the Fundação para a Ciência e Tecnologia (FCT) of Portugal. The APC was funded by COFAC, which is the host institution responsible for the management of Lusófona University where this study was conducted.

REFERENCES

Brocher, A., and Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology* 53, 1823–1835. doi: 10.1111/psyp.12770

Brocher, A., and Graf, T. (2017). Response: commentary: pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Front. Psychol.* 8:539. doi: 10.3389/fpsyg.2017.00539

Carvalho, J., and Rosa, P. J. (2020). Gender differences in the emotional response and subjective sexual arousal toward non-consensual sexual intercourse: a pupillometric study. *J. Sex. Med.* 17, 1865–1874. doi: 10.1016/j.jsxm.2020.06.018

Coney, J., and MacDonald, S. (1988). The effect of retention interval upon hemispheric processes in recognition memory. *Neuropsychologia* 26, 287–295. doi: 10.1016/0028-3932(88)90081-4

Cosme, G., Rosa, P. J., Lima, C. F., Tavares, V., Scott, S., Chen, S., et al. (2021). Pupil dilation reflects the authenticity of received nonverbal vocalizations. *Sci. Rep.* 11, 1–14. doi: 10.1038/s41598-021-83070-x

Federmeier, K. D., and Benjamin, A. S. (2005). Hemispheric Asymmetries in the time course of recognition memory. *Psychon. Bull. Rev.* 12, 993–998. doi: 10.3758/bf03206434

Granholm, E., and Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *Int. J. Psychophysiol.* 52, 1–6. doi: 10.1016/j.ijpsycho.2003.12.001

Heaver, B., and Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory* 19, 398–405. doi: 10.1080/09658211.2011.575788

Hess, E. H., and Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science* 132, 349–350. doi: 10.1126/science.132.3423.349

Joshi, S., Li, Y., Kalwani, R. M., and Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* 89, 221–234. doi: 10.1016/j.neuron.2015.11.028

Juvrud, J., Gredebäck, G., Åhs, F., Lerin, N., Nyström, P., Kastrati, G., et al. (2018). The immersive virtual reality lab: possibilities for remote experimental manipulations of autonomic activity on a large scale. *Front. Neurosci.* 12:305. doi: 10.3389/fnins.2018.00305

Kafkas, A., and Montaldi, D. (2012). Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. *Neuropsychologia* 50, 3080–3093. doi: 10.1016/j.neuropsychologia.2012.08.001

Kafkas, A., and Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology* 52, 1305–1316. doi: 10.1111/psyp.12471

Kafkas, A., and Montaldi, D. (2017). Commentary: pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Front. Psychol.* 8:277. doi: 10.3389/fpsyg.2017.00277

Kahneman, D., Onuska, L., and Wolman, R. E. (1968). Effects of grouping on the pupillary response in a short-term memory task. *Q. J. Exp. Psychol.* 20, 309–311. doi: 10.1080/14640746808400168

Kucewicz, M. T., Dolezal, J., Kremen, V., Berry, B. M., Miller, L. R., Magee, A. L., et al. (2018). Pupil size reflects successful encoding and recall of memory in humans. *Sci. Rep.* 8:4949. doi: 10.1038/s41598-018-23197-6

Lewandowska, K., Gągol, A., Sikora-Wachowicz, B., Marek, T., and Fąfrowicz, M. (2019). Saying "yes" when you want to say "no" - pupil dilation reflects

evidence accumulation in a visual working memory recognition task. *Int. J. Psychophysiol.* 139, 18–32. doi: 10.1016/j.ijpsycho.2019.03.001

Magliero, A. (1983). Pupil dilations following pairs of identical and related to be remembered words. *Mem. Cogn.* 11, 609–615.

Marques, J. F., Fonseca, F. L., Morais, A. S., and Pinto, I. A. (2007). Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables. *Behav. Res. Methods* 39, 439–444. doi: 10.3758/bf03193013

Naber, M., Frässle, S., Rutishauser, U., and Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *J. Vis.* 13:11. doi: 10.1167/13.2.11

Otero, S. C., Weekes, B. S., and Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology* 48, 1346–1353. doi: 10.1111/j.14698986.2011.01217.x

Papesh, M. H., Goldinger, S. D., and Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *Int. J. Psychophysiol.* 83, 56–64. doi: 10.1016/j.ijpsycho.2011.10.002

Rosa, P. J., Esteves, F., and Arriaga, P. (2015). Beyond traditional clinical measurements for screening fears and phobias. *IEEE Trans. Instrument. Meas.* 64, 3396–3404. doi: 10.1109/TIM.2015.2450292

Shepard, R. N., and Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *J. Exp. Psychol.* 62, 302–309. doi: 10.1037/h0048606

Sirois, S., and Brisson, J. (2014). Pupillometry. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 679–692. doi: 10.1002/wcs.1323

Steinhauer, S. R., Siegle, G. J., Condray, R., and Pless, M. (2004). Sympathetic and parasympathetic innervations of pupillary dilation during sustained processing. *Int. J. Psychophysiol.* 52, 77–86. doi: 10.1016/j.ijpsycho.2003.12.005

Unsworth, N., and Robison, M. K. (2018). Tracking working memory maintenance with pupillometry. *Attent. Percept. Psychophys.* 80, 461–484. doi: 10.3758/s13414-017-1455-x

van der Wel, P., and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: a review. *Psychon. Bull. Rev.* 25, 2005–2015. doi: 10.3758/s13423-018-1432-y

van Rijn, H., Dalenberg, J. R., Borst, J. P., and Sprenger, S. A. (2012). Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS One* 7:e51134. doi: 10.1371/journal.pone.0051134

Võ, M. L. H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., et al. (2008). The coupling of emotion and cognition in the eye: introducing the pupil old/new effect. *Psychophysiology* 45, 130–140. doi: 10.1111/j.14698986.2007.00606.x

Wang, C. A., and Munoz, D. P. (2015). A circuit for pupil orienting responses: implications for cognitive modulation of pupil size. *Curr. Opin. Neurobiol.* 33, 134–140. doi: 10.1016/j.conb.2015.03.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fpsyg.2021.748539

"Blue Sky Effect": Contextual Influences on Pupil Size During Naturalistic Visual Search

Steven M. Thurman[1], Russell A. Cohen Hoffing[1], Anna Madison[1], Anthony J. Ries[1], Stephen M. Gordon[2] and Jonathan Touryan[1]*

[1]US DEVCOM Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, MD, United States, [2]DCS Corporation (United States), Alexandria, VA, United States

Pupil size is influenced by cognitive and non-cognitive factors. One of the strongest modulators of pupil size is scene luminance, which complicates studies of cognitive pupillometry in environments with complex patterns of visual stimulation. To help understand how dynamic visual scene statistics influence pupil size during an active visual search task in a visually rich 3D virtual environment (VE), we analyzed the correlation between pupil size and intensity changes of image pixels in the red, green, and blue (RGB) channels within a large window (~14 degrees) surrounding the gaze position over time. Overall, blue and green channels had a stronger influence on pupil size than the red channel. The correlation maps were not consistent with the hypothesis of a foveal bias for luminance, instead revealing a significant contextual effect, whereby pixels above the gaze point in the green/blue channels had a disproportionate impact on pupil size. We hypothesized this differential sensitivity of pupil responsiveness to blue light from above as a "blue sky effect," and confirmed this finding with a follow-on experiment with a controlled laboratory task. Pupillary constrictions were significantly stronger when blue was presented above fixation (paired with luminance-matched gray on bottom) compared to below fixation. This effect was specific for the blue color channel and this stimulus orientation. These results highlight the differential sensitivity of pupillary responses to scene statistics in studies or applications that involve complex visual environments and suggest blue light as a predominant factor influencing pupil size.

Keywords: pupil size, luminance, pupillary light response, pupillometry, active visual search, virtual environment

OPEN ACCESS

Edited by:
Natasha Sigala,
Brighton and Sussex Medical School, United Kingdom

Reviewed by:
Rémy Allard,
Université de Montréal, Canada
Bruno Laeng,
University of Oslo, Norway

****Correspondence:***
Steven Thurman
steven.m.thurman3.civ@army.mil

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *29 July 2021*
Accepted: *16 November 2021*
Published: *21 December 2021*

Citation:
Thurman SM, Hoffing RAC, Madison A, Ries AJ, Gordon SM and Touryan J (2021) "Blue Sky Effect": Contextual Influences on Pupil Size During Naturalistic Visual Search. Front. Psychol. 12:748539. doi: 10.3389/fpsyg.2021.748539

INTRODUCTION

In classic cognitive pupillometry studies, it has been critical to equate luminance across stimuli and/or experimental conditions to isolate cognitive influences on pupil size and ensure that results are not driven by confounds due to variable luminance. While this careful control of luminance has led to a great deal of knowledge about the relationship between pupil size and cognition (Aston-Jones and Cohen, 2005; Sirois and Brisson, 2014; Mathôt, 2018; Hoffing et al., 2020; Joshi and Gold, 2020), it also limits the ability to generalize research findings to contexts in which luminance cannot be controlled. For example, it would be unfeasible to control for luminance in real-world tasks and would impede the study of naturalistic gaze

behavior in experiments using complex stimuli, to replicate the spatial variation in local contrast and luminance that occurs in the natural world (Frazor and Geisler, 2006).

The pupillary light response (PLR) represents a predictable, ballistic change in pupil size whenever there is a sudden change in luminance (Korn and Bach, 2016). As a rule, the pupil constricts and reduces in size whenever there is a sufficient increase in brightness and it dilates whenever there is a sufficient decrease in brightness, albeit more slowly than constrictions. The PLR is also a characteristically sluggish response that reaches its peak between 500 and 1,000 ms after a change in luminance (Mathôt, 2018), and only gradually returns to pre-stimulus baseline after several seconds. Gaps remain, however, in the understanding of how the pupil responds to light because it is still unclear how output from rods, cones, and intrinsically photosensitive retinal ganglion cells (ipRGC) are integrated to drive pupil size changes. For example, evidence from individuals with nonfunctioning rod and cone photoreceptors (Czeisler et al., 1995; Lockley et al., 1997) and transgenic mice without photosensitive RGCs (Lucas et al., 2001, 2003) suggest that both types of retinal cells contribute to the PLR; however, much less is understood about their relative contributions to the PLR, especially in uncontrolled settings with naturalistic and dynamic visual scenes.

It is well known that each type of photoreceptor has a different spectral sensitivity (Stockman et al., 1993; Do and Yau, 2010; Neitz and Neitz, 2011) and that these light-sensitive cells are non-uniformly distributed across the retina (Curcio et al., 1991; McDougal and Gamlin, 2010; Lee et al., 2017). Recent work leveraging the silent substitution method, which can selectively modulate the excitation of ipRGCs, rods, and the three cones separately (or combined), suggests that color signals influence the pupil response differently (Barrionuevo et al., 2014; Barrionuevo and Cao, 2016). For example, Bonmati-Carrion et al. (2018) found that monochromatic and combined monochromatic light had a differential influence on the strength of the PLR depending on the wavelength. Specifically, blue light (479 nm) resulted in a significantly faster velocity of constriction than purple (437 nm) or red (627 nm) light. Further complicating the matter, the PLR can also be modulated by contextual information, such as the expectation of a luminance increase, where participants showed increased PLRs to brightness illusions, such as the Helmholtz-Kohlrausch effect (Laeng and Endestad, 2012; Wood, 2012; Zavagno et al., 2017; Suzuki et al., 2019) and pictures of the sun (Naber and Nakayama, 2013; Castellotti et al., 2020), despite stimuli being equiluminant. By furthering our understanding of how dynamic visual scene statistics, such as luminance, spectral content (i.e., color), and context influence pupil size, it may be possible to better account for their contribution to the pupillary signal and improve estimation of residual cognitive-based effects on pupil size.

The present study aims to further our understanding of pupillary dynamics in a visually rich environment that involves an unconstrained navigation and visual search task in a 3D virtual environment (VE). Specifically, we focus on understanding how visual patterns modulate pupil size, where luminance changes dynamically over time even while behaviors and cognitive processes may also be concomitantly influencing pupil size. We investigated the influence of the spatial location of luminance in relation to the fovea as well as the spectral wavelength on pupil size changes. We hypothesized that the influence of luminance on pupil size would be greatest for pixels near the fovea and would reduce with eccentricity in a radial manner. This hypothesis is consistent with previous work indicating that the strength of the PLR is reduced as a function of eccentricity (Crawford and Parsons, 1936; Legras et al., 2018; Hu et al., 2020), which may be attributed to the diminishing distribution of photoreceptors farther away from the fovea. We also hypothesized that the relationship between luminance and pupil size would vary by wavelength, consistent with prior work indicating that blue colored light is perceived as being brighter (Suzuki et al., 2019) and can have a distinct influence on the PLR (Bonmati-Carrion et al., 2016). To investigate both hypotheses, we examined correlations between pupil size and intensities in the red, green, and blue (RGB) color channels derived from the sequence of images seen on the screen throughout the task. We computed pixel-wise correlation maps to visualize the correlation between pupil size and every pixel in a broad window (approximately 14 degrees visual angle) surrounding gaze position.

Foreshadowing our results from Experiment 1, the correlation maps surprisingly revealed that pixels closest to fixation actually varied the least with pupil size, contrary to our hypothesis of a foveal bias. The maps instead uncovered a distinct spatial pattern to indicate a significant contextual effect in which blue pixels, specifically located above the gaze position, had a disproportionate influence on pupil size. These results were interpreted to be related to a blue light from above or "blue sky effect," reasoning that from an ecological perspective it would make sense for the brain to anticipate a brightness change whenever there is a visual pattern resembling a blue sky overhead due to its association with daytime and sunlight (Laeng and Endestad, 2012; Naber and Nakayama, 2013; Castellotti et al., 2020). In Experiment 2, we performed a controlled laboratory experiment that paired a gray patch with luminance-matched red, green, and blue patches located either on the top, bottom, left, or right relative to the control (gray). Results of this follow-on study confirmed our hypothesis, demonstrating a significant and highly specific "blue sky effect" on the PLR.

Participants

Thirty-eight subjects were recruited from the Los Angeles area to participate in this study (Enders et al., 2021). Four subjects were missing a majority of eye-tracking data due to miscalibration or some technical error and were not included in this analysis, leaving a final sample of 34 subjects for this report (12 females, 22 males, age range = 19–64 years, mean = 39.5 ± 14.6 years). All subjects were at least 18 years of age or older and able to speak, read, and write English. All subjects signed an Institutional Review Board approved informed consent form prior to participation (ARL 19–122) and completed a web-based pre-screen questionnaire containing eligibility, demographic, and game-use questions. All subjects had normal hearing and

reported normal or corrected-to-normal visual acuity and color vision. Additional visual acuity screening was conducted in-lab to ensure better than 20/40 vision using a standard Snellen Chart. Subjects were asked to read the 20/40 line of the Snellen chart and were allowed to participate if they made one mistake or less (the clinical Snellen test allows patients to make up to two mistakes on a line to be classified as that level of visual acuity). Normal color vision was assessed with a 14-plate Ishihara color test. Any subject who did not pass the entire screening process was not included in the study.

Task and Procedure

Subjects completed demographic and survey questionnaires while being fit with an EEG cap prior to entering a whisper room (WhisperRoom Inc. MDL 4284 E) to undergo eye-tracking calibration and completed additional questionnaires pre- and post-tasks. The whisper room is a sound- and light-controlled chamber, where the only ambient lighting was provided by the computer screen. During the experimental session, subjects participated in four separate tasks including (i) classic rapid serial visual presentation (RSVP) target detection task (20 min), (ii) free viewing task to familiarize subjects with navigating the virtual environment (up to 12 min), (iii) the main free viewing visual search and navigation task (up to 20 min), and (iv) memory recall task (up to 15 min). All tasks were run with custom software using Unity 3D (Unity Technologies). Further documentation of the task and experimental design is described in previously published work (Enders et al., 2021). However, only results from pupillometry during the visual search and navigation task (iii) are described here.

In the visual search and navigation task, subjects were asked to freely navigate a virtual environment with the goal of searching for, and mentally counting, target objects from one of four categories that was randomly assigned to them (i.e., Aircraft, Motorcycle, Humvee, or Furniture). Subjects started at the same position in the virtual environment and all of the possible targets (15 total for each condition) were evenly distributed throughout the environment by experimenters to control the density of objects in each area (**Figure 1A**). Subjects had up to 20 min to identify and mentally keep count of the number of targets encountered using w/a/s/d keys for movement through the environment, and the mouse to control camera orientation to change heading direction and simulate head rotations. The task was performed on a computer monitor with a resolution of 1,920 × 1,080 pixels. Subjects were seated in a chair without a head or chin restraint and were asked to limit chair and body movements throughout the task. We used distance estimates from the eye-tracking system to confirm that subjects complied with this instruction. We found that subjects were positioned on average 62.1 cm from the screen (range = 55.0–78.0 cm) with a mean SD of 1.3 cm (range = 0.3–3.3 cm) over time.

About 8 min into the session an auditory Math Task was administered in which subjects were instructed to remember and report the sum of the numbers. Data collected during the math task, corresponding to 2.0 s (+0.47 s) on average, was cut from the time series data and omitted from analysis in order to reduce confounds associated with cognitive load and multi-tasking. Therefore, the main task during which we analyzed data in this report consisted of only the active visual search task, which had a mean duration of 10.4 min (+2.53 min). In this report, we leveraged this unique data set and the capability of replaying the entire set of visual scenes experienced by each subject to examine the relationship between pupil size and dynamic scene statistics irrespective of task-related cognition and behavior.

Luminance Measurements

Screen luminance measurements were obtained with a SpectraScan Spectroradiometer PR-745. We wrote a program in Matlab (2019a, The MathWorks, Natick, MA) using the Psychophysics Toolbox 3.0.16 (Brainard, 1997; Pelli, 1997) to step through each of the 8-bit RGB color channels from 0 to 255 in increments of 3, displaying the color on the full screen until the spectroradiometer collected luminance measurements in units of cd/m^2. When recording with the spectroradiometer, we matched the experimental setup to when subjects performed the task, and the distance of the device to the computer screen was positioned to match the average subject eye height and distance (62 cm) from the monitor. We fit the screen luminance data with an exponential function to estimate the best gamma parameter for transforming RGB color space into luminance space. Each color channel was best-fit by a slightly different gamma value (red gamma = 2.24, green gamma = 2.23, and blue gamma = 2.22). Prior to the experiment, we did not gamma correct the monitor by linearizing the color lookup table; instead, we applied this transformation *post hoc* to pixel intensities in each color channel, converting the images from 8-bit (0–255) color space to luminance space as a first step prior to subsequent analyses. We will use the term RGB luminance to reference this transformed image data.

Eye-Tracking Data Collection

Binocular eye-tracking data (300 Hz) were collected with a Tobii Pro Spectrum mounted on the bottom of the computer monitor. Prior to the main task, we used a standard five-point calibration procedure to ensure proper calibration of the eye-tracking system. The Tobii Pro Spectrum recorded binocular estimates of pupil size and gaze position, while eye-tracking data were synchronized with game state (i.e., positions of players and objects in the Unity environment), keyboard, and mouse data using the Lab Streaming Layer (LSL) protocol (Delorme et al., 2011; Preuschoff et al., 2011; Kothe and Makeig, 2013; Kothe, 2014). The Tobii Pro Spectrum is reported to have an average binocular accuracy of 0.3° and binocular precision (root mean square) of 0.07° (Tobii Pro, 2018).

Data Preprocessing

Post hoc data analysis involved generating the sequence of full-screen "snapshots" to replay the sequence of visual stimulation on the screen as subjects freely explored the virtual environment (**Figure 1B**). We generated the snapshot associated with every 10th frame for an effective temporal resolution of 12 Hz (screen

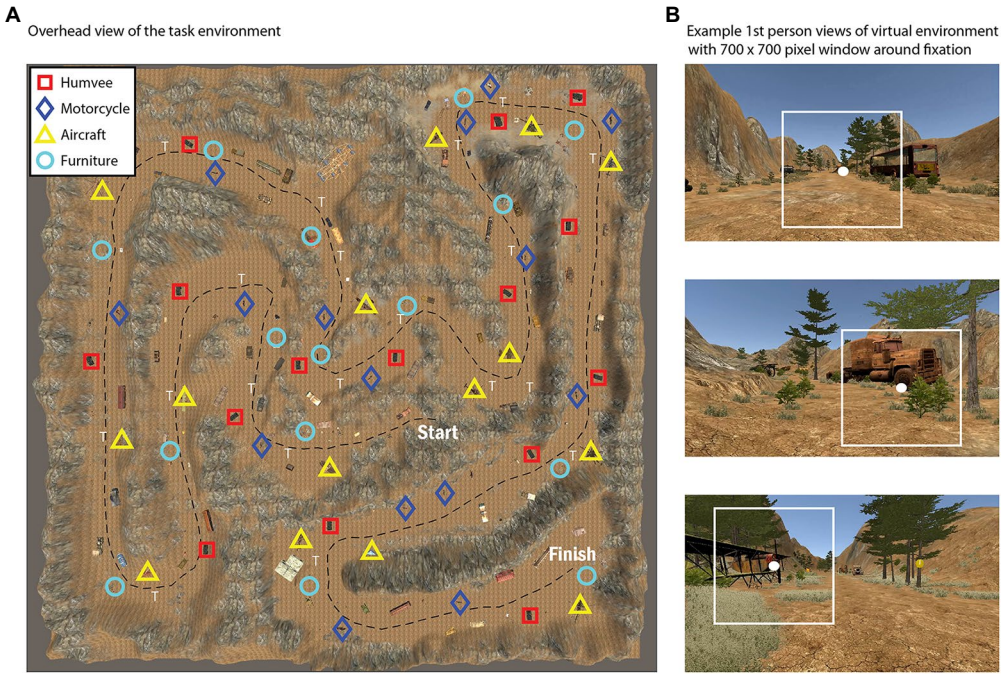


FIGURE 1 | Overhead view of the task environment during the active visual search and navigation task **(A)**, with symbols representing the location of various target types (Humvee, Motorcycle, Aircraft, or Furniture). For further illustration, **(B)** shows selected snapshots of the first person view during the task with a 700 × 700 pixel box surrounding fixation.

refresh rate was 120 Hz), exporting the images in .png format. To save disk space, the sequence of images was then compressed in Matlab using the built-in MPEG video encoder, reducing the file size footprint by about 100x from ~100 gb for all .png images to a video of ~1 gb per subject. To ensure that video encoding did not introduce significant artifacts to the image data, we compared RGB pixel values of the original .png images to the compressed video frames and found that over 99% of pixel intensity differences were within a range of +12, a criterion that corresponds to 5% of the entire 8-bit color spectrum (0–255).

We used gaze position data output from the Tobii Pro Spectrum to extract local image statistics in relation to gaze position over time. Signal loss due to blinks and dropouts in the gaze position data corresponded to 13.9% (STD = 11%) of the data overall. The range of missing data due to signal loss across subjects was 2.4–41.1%. To investigate whether this data loss influenced the reported pattern of results, we analyzed data from the subset of 29 subjects that had less than 25% data loss (omitting the five subjects with greater than 25% data loss) and the overall pattern and statistical significance of group results were not affected whether including or excluding these subjects. Thus, these subjects were included in the presented results.

Instead of averaging pupil size data from the two eyes, which can introduce artifacts (e.g., abrupt discontinuities) when data are missing from one eye but not the other (due to baseline differences between the two eyes), we selected *a priori* to analyze the eye with the least amount of missing data (e.g., due to blinks and signal dropout due to eye/head rotations). Estimates of pupil size from commercial eye trackers can be noisy due to challenges in fitting the pupil region with an ellipse in the presence of eye lashes, partial eye closures, squinting, eye rotations, and other factors. These factors can sometimes introduce artifacts that appear as very abrupt and large changes in pupil size from one sample to another that are physiologically unrealistic. To reduce these artifacts in the pupil time series, we used an iterative velocity-based approach that examined the overall distribution of velocities over time and first identified all values that were greater than +2 SDs away from the mean and replaced them with not a number (NaN). It then used a more stringent criterion on the second iteration to remove values greater than +2.5 SDs from the mean to remove any remaining large outliers. These missing data points were then filled-in using linear interpolation of nearby data points.

A blink causes the eyelid to momentarily occlude the eye and causes a brief signal dropout because the eye is no longer visible to the tracker. Blinks were defined by short sequences of signal dropout that ranged from 50 to 500 ms of contiguously missing data. Missing data due to blinking were linearly interpolated using the best practice of also removing several data points (up to 50 ms) pre-blink and post-blink to remove potential artifacts due to partial eye closure surrounding each blink. We used the procedure published in the PRET toolbox (Denison et al., 2020), which is based on the technique published by Mathôt (2013). This technique first smooths the data with an 11 ms Hanning window and uses a velocity-based threshold

to detect the onset and offset of each blink within a time window of +50 ms surrounding the epoch of missing data. The interpolated pupil size data and raw gaze position data were then downsampled to the same temporal resolution as the snapshot images (12 Hz) for subsequent analyses.

Data Analysis

We computed correlation maps representing the Pearson correlation coefficient between pupil size and every pixel within a large region (700 × 700 pixels) centered on gaze position (+350 pixels, or approximately 7.6 deg. in each direction) derived from the snapshot images. To account for the expected time delay between changes in brightness/darkness and changes in pupil size due to the well documented sluggishness of the PLR (Mathôt, 2018; Denison et al., 2020), we used an empirical approach to estimate an appropriate time delay by cross-correlating pupil size with the RGB time series, and examining the peak temporal offset that maximized the correlation. We repeated this for every color channel (three channels) and subject (34 subjects), which resulted in a normal distribution of time lags with a mean = 469.4 + 140 ms and median = 500 ms (i.e., six frames at 12 Hz). This value is within the range of expectations according to prior literature (Ellis, 1981). Based on these results, we decided to fix the temporal offset at 500 ms for all subsequent analyses prior to computing correlation coefficients by shifting the pupil time series forward by six frames (500 ms) relative to RGB luminance. Of note, we performed follow-up analyses to investigate the influence of temporal offsets by varying the length of the temporal offset between 0 (no pupil lag) and 1,000 ms and found that the overall pattern of results was highly robust to the choice of lag value.

For computational tractability with these rather large image stacks, we downsampled the images by a factor of 10, from 700 × 700 to 70 × 70 pixels, where each pixel in the correlation map was associated with the average intensity of a 10 × 10 block of pixels (approximately 0.2 deg.) with reference to the original resolution. When the gaze position was too close to the edge of the screen (i.e., within 350 pixels), we replaced pixels in the square window that would have fallen off-screen as NaNs in the image stack so they would not be incorporated into the correlation analysis. The total amount of missing image data (NaN values) due to fixations near the edge of the screen was 6.47%, and the percentage reached a maximum of 18% for pixels at the very top of the images. A figure representing the proportion of NaN values across space resulting from this procedure is shown in **Supplementary File 1**.

Because correlation coefficients are distributed non-normally, we performed a Fisher Z transform prior to computing group-level statistics (Dunn and Clark, 1969; Lenhard and Lenhard, 2014). We examined the consistency of spatial patterns in the correlation maps at the group level by averaging Fisher Z-transformed correlation maps across subjects and performing one-sample *t*-tests on a pixel-by-pixel basis comparing the distribution of values to the null hypothesis of 0 (no correlation). Due to the large number of pixels (4,900) included in this analysis, we used the procedure from Benjamini and Hochberg (1995) commonly used in fMRI to control the false discovery rate (FDR). We set a conservative criterion of FDR = 0.01 for analysis of pixels in the correlation maps.

To compare between each pair of color channels directly, we computed dependent-samples correlations (Williams, 1959; Steiger, 1980), as recommended by Steiger (1980), which tested the hypothesis that a pair of color channels was equally correlated with pupil size using a t-distribution. This statistical test is appropriate in such cases when two repeated-measures variables (i.e., two color channels) derived from the same individual are correlated with a third variable (i.e., pupil size). High *t*-values provide evidence to reject the null hypothesis and indicate a significant difference in the strength of the correlation between pupil size and one color channel vs. another color channel.

Due to the inverse relationship between luminance and pupil size, where an increase in brightness causes a pupillary constriction and a decrease causes dilation, we expected a negative relationship between pupil size and RGB luminance and therefore that pixels with a stronger influence on pupil size would show larger negative correlations. Similar to other research areas, for example, that use reverse correlation to construct maps that reveal the influence of patterns of stimulus information on behavior (Gosselin and Schyns, 2001; Thurman et al., 2010; Thurman and Lu, 2013), we expected that the correlation map technique used here would be powerful enough to uncover fine-scale spatial patterns to characterize the influence of individual pixel intensities on pupil size. Consistent with research showing an effect of eccentricity such that the strength of the PLR is strongest for stimuli presented near the fovea and is systematically weaker for stimuli presented farther away, we hypothesized a foveal bias in the correlation maps (**Figure 2**) indicating that the visual system would pool luminance information from a focal region surrounding gaze position to modulate pupil size. An alternative possibility, however, would be that pixels were weighted in an anisometric pattern reflecting a broader contextual influence of luminance information on pupil size. In the absence of a specific theoretical prediction, an unexpected pattern, such as this would be interesting and informative, but would require *post hoc* analysis for interpretation.

Results

Behavior

The analyses presented in this paper were focused on characterizing the relationship between pupil size and RGB scene statistics over the course of a visual search task in a virtual environment. Though our analyses are agnostic to behavioral performance in this task *per se*, it is relevant to report whether subjects were on-task and successful in reporting the correct number of target objects to be identified, and the correct answers to the mental math questions (see Task and Procedure). In terms of identifying and recollecting the number of targets (there were 15 target objects total for each condition), 25% of subjects reported exactly 15 targets but 71.9% of subjects did report at least 14 targets (range 5–32 targets). The variance could have been due to a misunderstanding by some subjects regarding which objects seen were supposed to be part of the

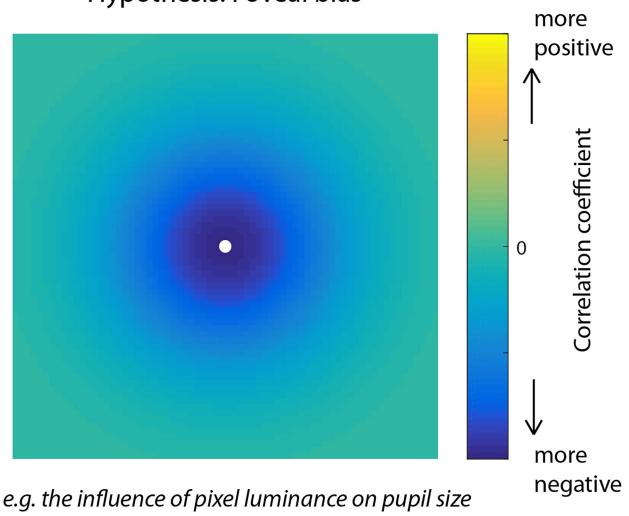


FIGURE 2 | Visual representation of the foveal bias hypothesis for the correlation map analysis of Experiment 1. Due to the density of photoreceptors and sensitivity of the pupillary light response (PLR) to foveal vs. peripheral light stimulation, we predicted that pixels nearest fixation would correlate the most negatively with pupil size over time, and pixels further away would correlate less. The predicted negative correlation is due to the inverse relationship between light intensity and pupil size in which increases in luminance induce constrictions of the pupil.

target class they were assigned (Humvees, Motorcycles, Aircraft, or Furniture). On average, subjects had a mean accuracy of 75.2% in reporting the correct sum for the mental math questions and 94.3% of subjects got at least one of the three math questions correct. The behavioral results indicate that a majority of subjects were on-task in performing the visual search and mental math tasks as instructed.

Correlation Maps

The correlation map analysis allowed us to explore how different regions (at the pixel level) surrounding gaze position contributed to pupil size fluctuations relative to other regions. Critically, it does not assume a particular spatial pattern underlying the relationship between color luminance and pupil size; rather, the correlation maps allow us to discover patterns in the data. **Figure 3** shows mean group-level correlation maps (left) for each color channel as well as thresholded statistical maps (t-scores; middle) highlighting pixels in which the distribution of correlation values between-subjects was significantly different from zero (FDR < 0.01). The subpanels in **Figure 3** (right) show correlation maps derived from each individual subject to help visualize consistency of results between-subjects. We observed that the spatial patterns of these individual correlation maps were not as consistent from subject to subject for the red channel (**Figure 3**, top row), but were much more consistent for the green and blue channels (**Figure 3**, middle and bottom rows). In particular, the green and blue channels revealed a systematic bias showing a much stronger negative correlation for pixels above fixation compared to pixels below fixation. To evaluate this apparent upper bias in the blue channel for each subject statistically, we performed a t-test comparing the distribution of correlations above fixation to those below and found that 33/34 subjects showed a statistically significant difference ($p < 0.05$). There was no such consistency in the red channel across subjects, indicating a blue-green specificity for this upper visual field bias.

Contrary to the foveal bias hypothesis that predicted pixels nearest to fixation would correlate the most with pupil size (**Figure 2**), the correlation maps did not reveal an isometric influence of pixels in a circular region around fixation, nor did it show a pattern consistent with a non-linear weighting of pixels as a function of the distance from fixation (e.g., a Gaussian blob). Instead, the correlation maps showed the opposite pattern in which pixels closer to fixation were relatively less correlated with pupil size as evidenced by the statistically non-significant (n.s.) regions nearest to fixation in the thresholded t-score maps (**Figure 3**, middle column). Inspecting the individual subject maps, this result appears to be driven by inconsistency between-subjects in the central region, in which some subjects actually showed a positive relationship between pixel luminance and pupil size (yellow areas of the individual maps) and other subjects showed the opposite effect or, in most cases, a weak relationship in the central region. These results provide evidence decidedly against the foveal bias hypothesis and strongly support that contextual information outside the fovea can significantly modulate pupil size.

In the analyses presented above, the correlation maps illustrated pixels that were significantly correlated with pupil size with reference to the null hypothesis of no correlation (e.g., $r = 0$), but do not indicate whether the correlation for one color channel was significantly greater than another color channel. To compare between pairs of color channels, we performed a comparison of correlations for dependent samples that takes into account the within-subjects correlation of two variables (i.e., color channels) with a third variable (i.e., pupil size) as well as the correlation between the two variables (Williams, 1959; Steiger, 1980). As shown in **Figure 4**, the comparison of Blue-Red (left) and Green-Red (middle) revealed that the correlation of pixels in the upper part of the visual field was significantly more negative for both Blue and Green in comparison to Red. The comparison of Blue-Green (right) further showed that blue was significantly more correlated with pupil size than green for a smaller subset of pixels at the very top of the map. There were no significant differences among the color channels for pixels in the lower part of the visual field. This result further demonstrates that information in the blue and green channels, specifically located above fixation, had a dominant and disproportionate influence on pupil size that was much stronger than the red channel throughout the active visual search task.

Green and blue showed a similar pattern of results in the correlation maps, in part, because they were strongly correlated over time, specifically for pixels located in the sky region. In an attempt to isolate the influence of green vs. blue, we performed a follow-on analysis that converted the images from 8-bit RGB image space to CIELAB space, which is an alternative representation

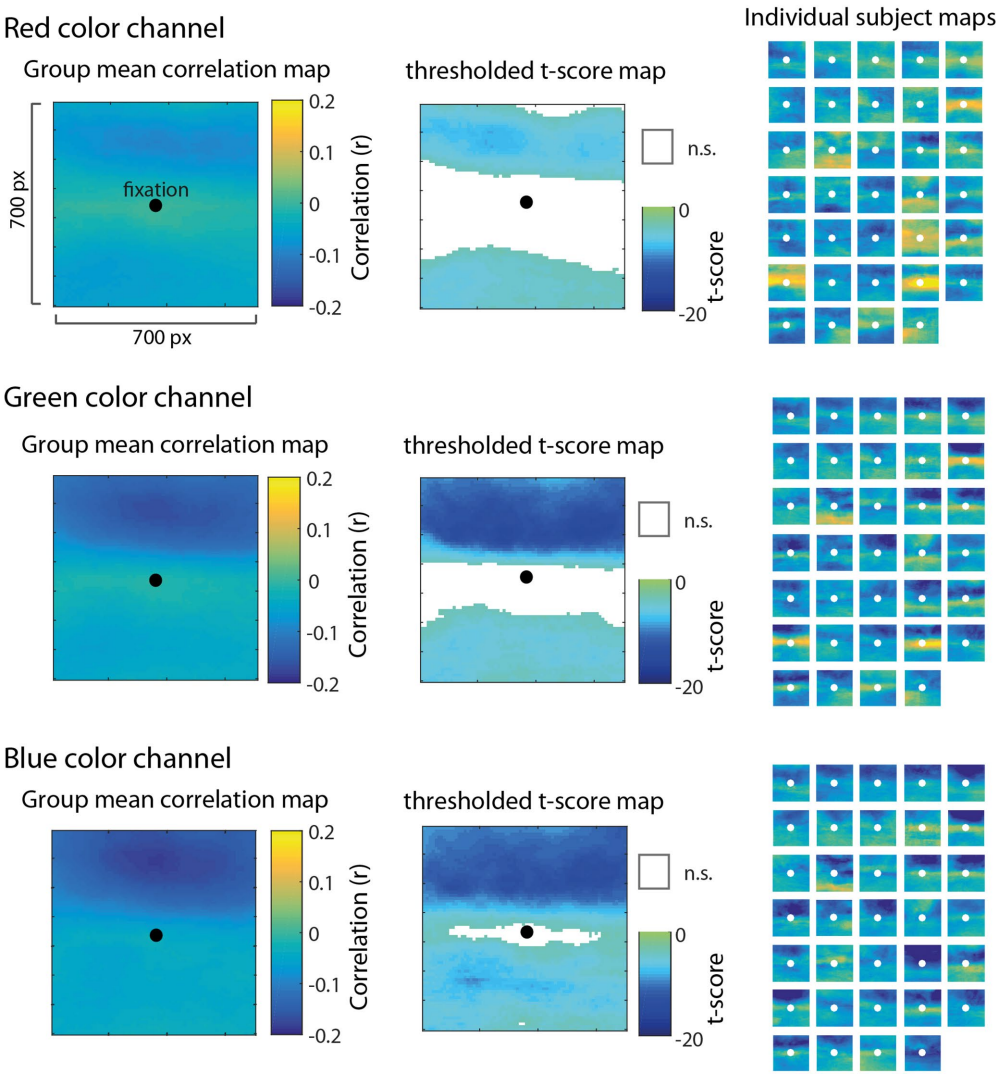


FIGURE 3 | Results of the correlation map analysis for the red (top row), green (middle row), and blue (bottom row) channels, illustrating the correlation between fluctuations in pupil size and pixel intensity in a 700 × 700 pixel square region surrounding fixation. Group mean correlation maps (left) are shown with a black circle as a reference to indicate the center of gaze. We computed the *t*-score on Fisher Z-transformed correlation values for each pixel to derive statistical maps (center) and applied a false discovery rate threshold (FDR = 0.01) with non-significant pixels (n.s.) represented as white. Individual maps for each of 34 subjects (right) are shown to visually inspect the consistency of results across subjects.

of an image in a three dimensional space that is more aligned with human perception. The three channels of CIELAB space include L* which is a representation of luminance on a scale of 0 (lowest luminance) to 100 (highest luminance), a* which is a representation of red-green chromaticity on a scale of −100 (more red) to 100 (more green), and b* which is a representation of blue-yellow chromaticity on a scale of −100 (more blue) to 100 (more yellow). We computed correlation maps using the same procedure as before, except in this case, we used pixel values represented in the three channels of L*a*b* space.

Results of this analysis are shown in **Figure 5**. The correlation map associated with the luminance channel (top row) showed a similar pattern to green and blue from the original analysis, in which the luminance of pixels above fixation had a statistically significant and disproportionate influence on pupil size. **Figure 5** (middle row) shows that information in the a* channel was not very strongly correlated with pupil size, indicating that chromaticity along the red-green dimension was not a predominant signal influencing pupil size. By contrast, **Figure 5** (bottom row) shows that information in the b* channel, representing chromaticity along the blue-yellow dimension, had a striking association with pupil size, particularly for pixels above fixation. The significant positive correlations in this map indicate that pupillary constrictions (reductions in pupil size) were associated strongly with increases in blue chromaticity (more negative values in b* space). This result provides additional clarity to the RGB results, and further evidence to suggest that there is a highly specific sensitivity of the PLR to visual patterns that indicate a blue sky is overhead.

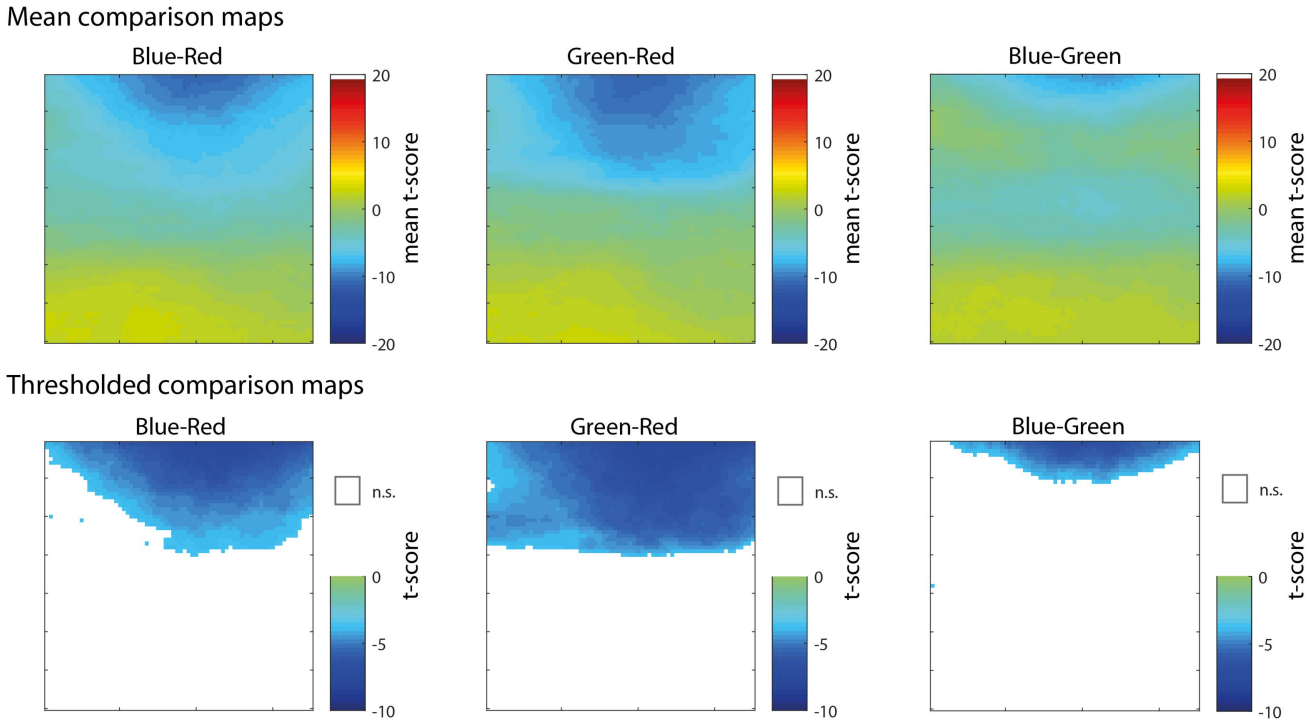


FIGURE 4 | Maps showing pixel-wise *t*-scores (top row) for the comparison of correlations for dependent samples, comparing blue-red channel (left), green-red channel (middle), and blue-green (right). Thresholded versions of the same maps (bottom row) showing significant pixels (FDR < 0.01) colored by *t*-score according to the color map and non-significant pixels represented as white.

EXPERIMENT 2

In Experiment 1, we found that there was a stronger negative correlation between pupil size and blue pixels located in the top portion of images relative to fixation. We hypothesized this result as being associated with a "blue sky effect," or a blue light from above effect (Suzuki et al., 2019). This hypothesis is motivated by three features of the visual system; first, that it is biased toward perceiving blue light as brighter; second, that it incorporates expectations (i.e., priors) of the structure of the environment such that visual patterns associated with sunlight from above are associated with an expectation of increased brightness; third, that the PLR has adapted to have increased sensitivity (indexed by a stronger PLR) to such patterns associated with sunlight from above (e.g., a blue sky).

To test whether the aforementioned effect can be explained by increased sensitivity to blue light above fixation, we designed a follow-up experiment that presented luminance-matched color stimuli (red, green, and blue) separated by either the horizontal or vertical meridian of the computer screen and paired with luminance-matched gray on the other side of the screen. We hypothesized that (i) blue would result in a larger constriction of pupil size due to a general bias or sensitivity to blue light, (ii) that blue light above would induce larger pupillary constrictions compared to blue light below, and (iii) that this effect would be specific for the blue channel and this orientation. These hypotheses are consistent with data presented in Experiment 1, and an ecological perspective that years of experience in the world with a blue sky (correlated with sun brightness) has adapted the system to anticipate or exaggerate the PLR specifically when blue is overhead.

Participants

Thirty subjects with reported normal vision participated in this study. Due to the fact that interpolation introduces significant distortions in the shape of the PLR, we excluded trials in which a blink occurred in the first 1,500 ms following stimulus onset. Accordingly, we removed subjects from analysis if they had too many trials discarded due to ill-timed blinks according to the following criterium: (1) they must have had at least one valid (non-blink) trial for each condition (12 total conditions) and (2) at least 50% of trials overall must have been valid (non-blink). In total, 15 subjects (five females, 10 males, mean = 20.3 + 3.46 years) met these criteria and were included in the analysis. All subjects signed an informed consent form approved by the Institutional Review Board (ARL 20–014) prior to participation and completed a demographics questionnaire. The experimental protocol and human subjects procedures were in compliance with the Declaration of Helsinki.

Task and Procedure

This data set was collected as a part of a larger data collection effort, where subjects first completed an attentional cueing task

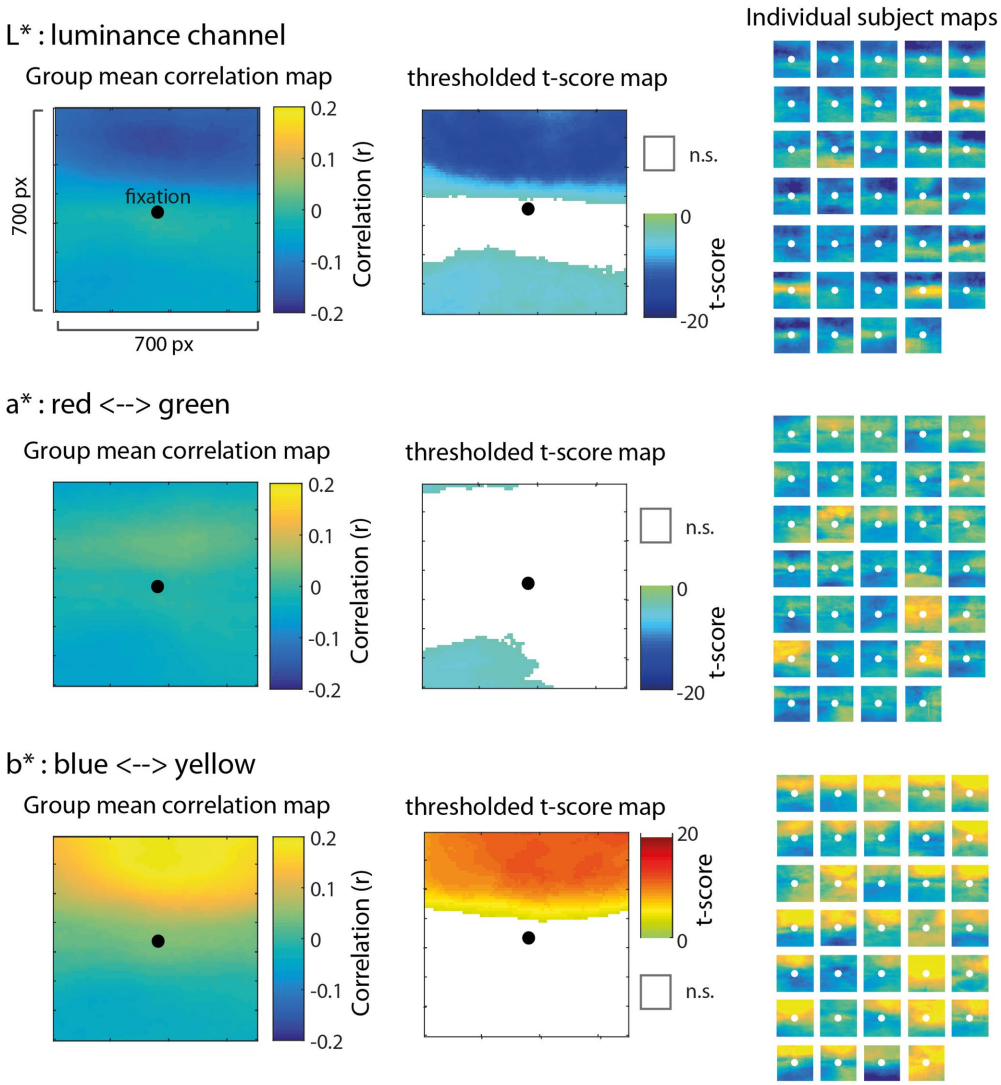


FIGURE 5 | Results of the correlation map analysis for images transformed to CIELAB space with L* (top row), a* (middle row), and b* (bottom row) channels, illustrating the correlation with pupil size of each of the three dimensions in a 700 × 700 pixel square region surrounding fixation (similar to **Figure 3**). In CIELAB space, L* represents luminance on a scale of 0–100, a* represents chromaticity along the red (−100) and green (+100) axis, and b* represents chromaticity along the blue (−100) and yellow (+100) axis. Group mean correlation maps (left) are shown with a black circle as a reference to indicate the center of gaze. We computed the *t*-score on Fisher *Z*-transformed correlation values for each pixel to derive statistical maps (center) and applied a false discovery rate threshold (FDR = 0.01) with non-significant pixels (n.s.) represented as white. The positive correlation for pixels above fixation in the b* channel (bottom) indicates that decreases in pupil size (e.g., light-induced pupillary constrictions) were strongly associated with more negative values in b* space (e.g., increases in blue chromaticity). Individual maps for each of 34 subjects (right) are shown to visually inspect the consistency of results across subjects.

prior to completing the light from above task. The study had a total duration of 40 min and the task (described below) had a duration of 7 min. Only data from this specific task is reported here. After realizing that several early subjects were blinking too often during the first 1,500 ms of each trial, we modified our instructions asking subjects explicitly to withhold blinks for the first several seconds following stimulus onset. This resulted in less blinks during the critical period of the PLR for subsequent subjects.

In this task, we split the computer screen along either the horizontal or vertical meridian, with one half of the screen colored gray and the other size colored with luminance-matched red, green, or blue. The order of stimulus presentation was pseudorandomized and counterbalanced by three color conditions (red, blue, and green), and four location conditions (top, bottom, left, or right) resulting in a total of 12 conditions (**Figure 6A**). A white fixation circle of 0.1 degrees was always present on the screen to help subjects maintain fixation. Each condition was repeated five times for a total of 60 stimulus presentations. Stimuli were presented for 300 ms followed by a black screen presented during the inter-stimulus-intervals (ISI) with a randomly jittered ISI between 3,000 and 5,000 ms (**Figure 6B**).

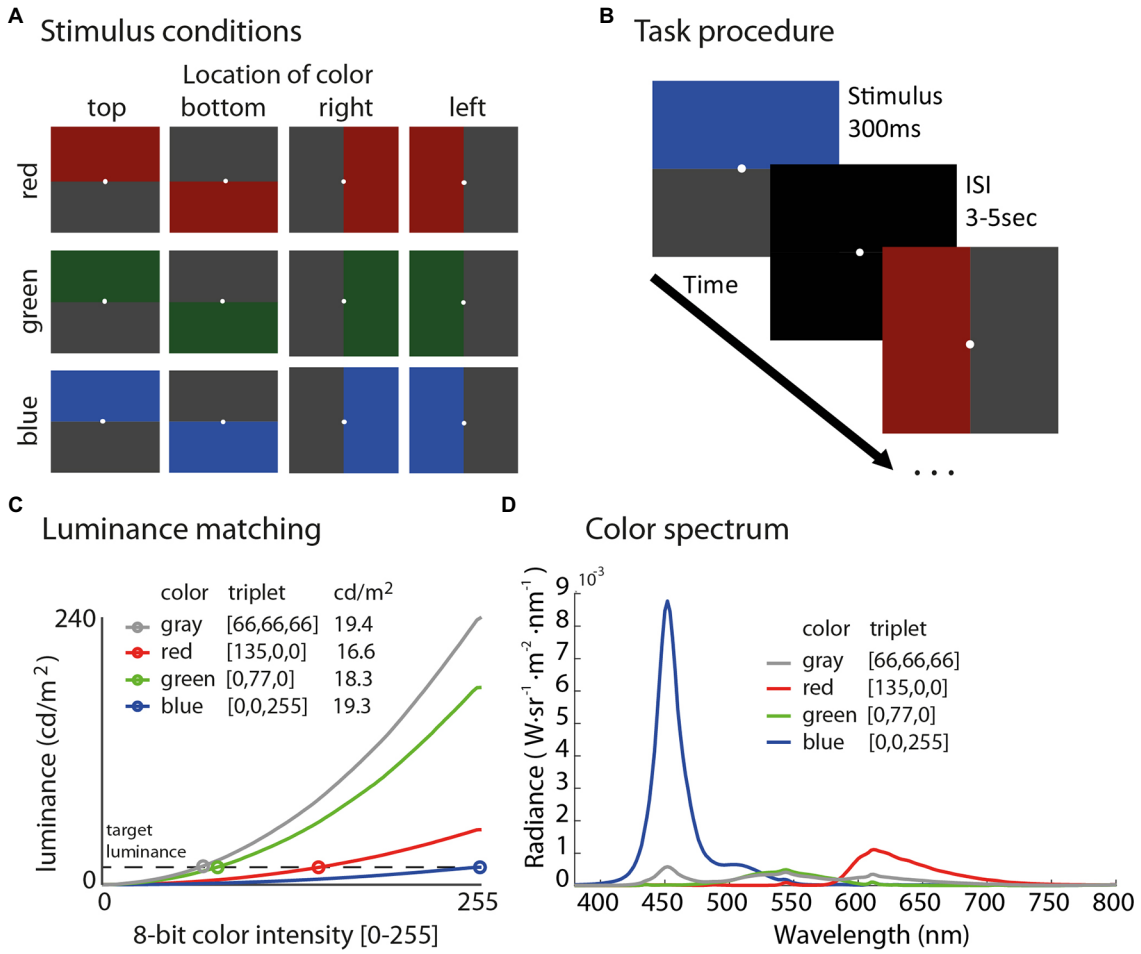


FIGURE 6 | Illustration of all 12 stimuli in the experiment **(A)** sorted by color in rows and location in columns. Schematic of the task procedure **(B)** shows that subjects fixated on a white dot and were presented each stimulus for 300 ms followed by an inter-stimulus interval (ISI) with a black screen for 3–5 s. Stimuli were luminance matched using measurements from a spectroradiometer **(C)** to evaluate the specific influence of color and location on the PLR, while controlling for overall luminance between conditions, which ranged from 16.6 to 19.4 cd/m^2. The color spectrum plot **(D)** shows radiance as a function of wavelength derived from the spectroradiometer for each color stimulus.

The luminance of the stimuli was measured using the same spectrophotometer (SpectraScan Spectroradiometer PR-745) and protocol as in Experiment 1.

As shown in **Figure 6C**, all colors fell within the range of 16.6–19.4 cd/m^2 [red (135,0,0) = 16.59 cd/m^2, green (0,77,0) = 18.23 cd/m^2, blue (0,0,255) = 19.28 cd/m^2, and gray (66,66,66) = 19.43 cd/m^2]. We chose this range of luminance on the basis of maximizing luminance in the blue channel to evoke the strongest possible PLRs; then, finding the RGB triplet in each color channel that was the best match to the maximum luminance of blue (19.28 cd/m^2) based on our measurements with the spectroradiometer. **Figure 6D** shows the color spectrum plot for each stimulus used in the experiment.

The task was programmed in Matlab (2014a, The MathWorks, Natick, MA) and the Psychophysics Toolbox (3.0.14) was used for stimulus presentation (Brainard, 1997; Pelli, 1997). Stimuli were presented on a 2,560 × 1,440 Acer XB271HU monitor with a 120 Hz refresh (Windows 7 64-bit, Nvidia GeForce GTX 660 video card, Intel i7-4770K 4-core 3.5 GHz CPU, 16 GB RAM).

Eye-Tracking Data Collection

Eye-tracking data were acquired in pupil-corneal reflection tracking mode (centroid pupil tracking) from the left eye, sampled at 1,000 Hz using the EyeLink 1,000 Plus eye tracker (SR Research, Ontario, Canada). Subjects were seated 75 cm from the monitor, while positioned in a chin rest to minimize head movements, and underwent a nine-point calibration procedure. Subjects recalibrated until an average validity error less than 1 deg. was obtained.

Data Preprocessing

Pupil size data output from the EyeLink device is recorded in arbitrary units, so we first normalized pupil size on each trial to percent signal change by subtracting the pre-stimulus baseline, and then dividing by the baseline and multiplying by 100. The baseline was defined as the mean pupil size in the interval of 500 ms prior to stimulus onsets. We then averaged all valid trials (trials without blinks) associated with each condition within-subjects for subsequent group-level analyses.

Data Analysis

Our analysis focused on characterizing and comparing the strength of the PLR, which is a ballistic constriction of the pupil initiated by the onset of a relatively brighter stimulus. Stimuli on each trial were always paired such that a colored stimulus (red, green, or blue) was presented with a luminance-matched gray (**Figure 6A**), so mean luminance across the entire screen (and between color conditions) was near constant (ranged from 16.6 to 19.4 cd/m^2). Therefore, any difference in the strength of the PLR would be the result of differential sensitivities related to color spectrum, visual field location (top, bottom, left, and right), and/or an interaction between these two factors.

We quantified the strength of the PLR by identifying the minimum value (PLRmin = peak pupillary constriction) in the time interval from 0 to 2 s post-stimulus onset. We ran a repeated-measures ANOVA on PLRmin values with two factors including color (red, green, and blue) and color location (top, bottom, left, and right) to evaluate main effects and interaction effects between these factors. We corrected any violations from the assumption of sphericity with the Greenhouse-Geisser correction. Planned comparisons were also performed to assess the strength of the PLR specifically for top – bottom and left – right conditions to derive a set of difference scores for each color channel. We evaluated one-sample *t*-tests on these difference scores to assess whether the distribution was significantly different from the null hypothesis of zero. A difference score of zero would indicate that the relative location of the color did not impact the strength of the PLR, whereas a score significantly different from zero would indicate an effect due to the location of color (top compared to bottom or left compared to right). To correct for multiple comparisons (six total, two difference scores by three colors), we report adjusted *p*-values for *t*-tests using the Benjamini and Hochberg (1995) procedure to control the false discovery rate (FDR = 0.05).

Results

Pupillary light response waveforms for each condition are shown in **Figure 7** (top), organized by color and orientation. The repeated-measures ANOVA revealed a significant main effect of color [$F(2,28) = 43.67$, $p < 0.001$, $\eta^2 = 0.56$] due to blue being associated with a much stronger PLR (mean = −44.3%, SE = 2.0%) by comparison to red (mean = −37.7%, SE = 2.0%) and green (mean = −37.2%, SE = 2.0%). The main effect of location was not significant [$F(3,43) = 0.76$, $p = 0.52$], but we did find a significant interaction effect between color and location [$F(6,84) = 2.72$, $p = 0.018$, $\eta^2 = 0.03$] indicating that the influence of color was also modulated by its location on the PLR.

To further probe this interaction effect, we conducted one-sample *t*-tests on the PLRmin difference scores for each color channel and orientation type (vertical or horizontal; **Figure 7**, bottom). Blue-top – blue-bottom was the only comparison that was significantly different from zero [$t(14) = -3.60$, $p = 0.003$, FDR < 0.05] indicating that the PLR was significantly greater specifically when blue was located on top compared to when blue was located on bottom. All other comparisons (red/top-red/bottom, red/left-red/right, green/top-green/bottom, green/left-green/right, and blue/left-blue/right) were non-significant (all values of $p > 0.05$, FDR > 0.05). This effect could not be explained by baseline differences in pupil size prior to the stimulus, as the 500 ms mean pre-stimulus baseline pupil size for blue-top was 913 a.u. (*SD* = 256 a.u.) and for blue-bottom was 926 a.u. (*SD* = 282 a.u.), and the *t*-test indicated a non-significant difference in baseline pupil size, $t(14) = 0.53$, $p = 0.6$. In fact, there was a non-significant difference in baseline pupil size for all comparison pairs ($p > 0.05$), ensuring that differences in baseline pupil size could not explain this pattern of results. This highly specific effect for blue was consistent with our hypotheses that the response of the pupil to blue light would be greater than red and green, and also that blue on top would induce a significantly larger PLR than blue on bottom. The specificity of this result is consistent with the hypothesized blue sky effect and provides a potential explanation for the significant contextual effect in Experiment 1, in which blue pixels above fixation had the strongest modulatory effect on pupil size.

DISCUSSION

In this study, we leveraged a unique data set in which subjects performed a navigation and active visual search task in a complex 3D virtual environment to examine the relationship between fluctuations in pupil size and dynamic visual scene statistics. Our primary goal was to better characterize how the pupil is influenced by rapid changes in complex luminance patterns associated with realistic scenes. We did this by correlating pupil size and pixel intensities in the RGB channels of the image in relation to the center of gaze as subjects actively performed the task. A primary motivation for this work is that while there are many published studies that have characterized the PLR in controlled laboratory experiments, most of these studies use simplistic stimuli and longer stimulus durations that do not match the complexity and high temporal rate of change of luminance in realistic visual scenes. It is unknown the extent to which findings from these studies will generalize to complex environments to enable effective use of cognitive pupillometry in less constrained task environments and real-world applications. We tested whether three findings from previous literature on the relationship between pupil size and luminance would generalize to a naturalistic task: (i) pupil size is more strongly influenced by luminance inputs near the fovea, (ii) pupillary responses to light are modulated by color spectrum effects (given equal luminance), with a specific sensitivity to blue light, and in Experiment 2, and (iii) that pupil size is more strongly influenced by blue light, specifically when it is located above fixation.

No Fovea Bias

We hypothesized a foveal bias in the correlation maps, consistent with research showing that stimuli nearer to fovea induce

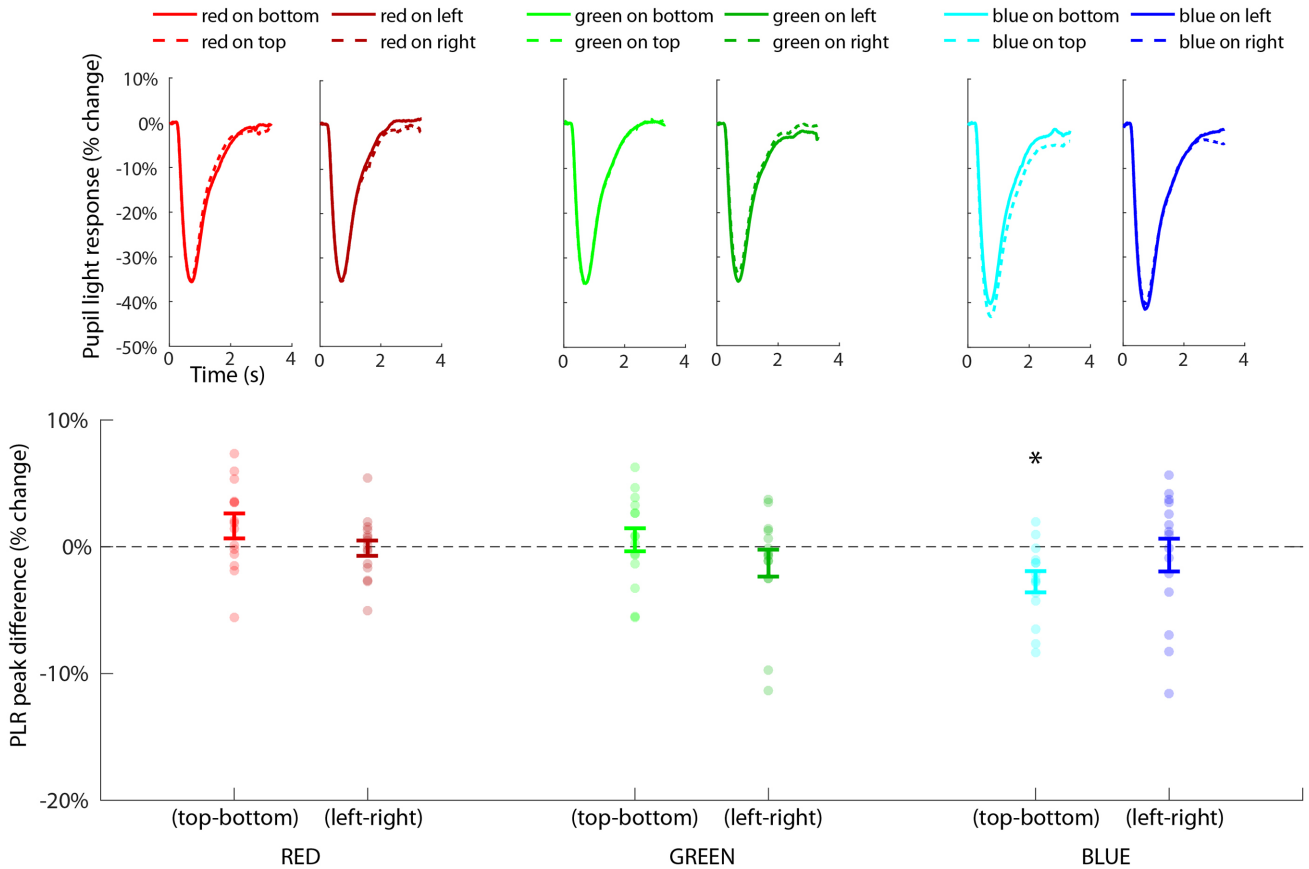


FIGURE 7 | Results of Experiment 2 showing the PLR for each condition (top), organized by color and orientation (i.e., horizontal and vertical) along the x-axis. Statistical tests were conducted on the distribution of PLRmin difference scores (bottom) across subjects (i.e., individual dots) for each planned comparison of each type of color/orientation combination. The only significant effect ($p < 0.02$, FDR < 0.05) was the comparison of blue on top minus blue on bottom (indicated with *) showing that blue on top induced a significantly larger PLR than blue on bottom.

stronger PLRs than stimuli in the periphery (Crawford and Parsons, 1936; Legras et al., 2018; Hu et al., 2020). This prediction is also related to the structure of the human visual system as the distribution of photoreceptors changes dramatically as a function of distance from the fovea, and there are significant differences in visual function between foveal, parafoveal, and the peripheral regions (Preuschoff et al., 2011; Strasburger et al., 2011). Previous research has shown that cone mediated pupil responses to photopic light (i.e., simple circular patches of light stimulus) are maximal up to 7 degrees of visual angle, and has reasoned that the decrease in response may be due to the drop off in cone density after 7 degrees (Legras et al., 2018; Kelbsch et al., 2019). Our data, however, did not support this hypothesis, instead revealing a robust and unpredicted pattern in the correlation maps. Pixels very close to gaze position actually tended to correlate the least strongly with pupil size. This pattern was most prominent for the red channel, in which all of the pixels within about 2–3 degrees from fixation were not significantly associated with pupil size, but it was also apparent in the blue and green channels.

Blue Sky Effect

The correlation maps instead showed evidence to support a strong contextual influence on pupil size associated with the presence of blue occurring above fixation in the environment, that we refer to as the "blue sky effect." Our finding that blue varies more strongly with pupil size compared with red or green (**Figure 4**) is consistent with previous research. For example, the Helmholtz-Kohlrausch effect indicates that color saturated light is perceived as being brighter than light that is less saturated at equiluminance, and that this effect is increased for particular wavelengths including blue (Wood, 2012; Suzuki et al., 2019). Likewise, eye-tracking studies find that a perceived increase in brightness coincides with pupillary changes (Sulutvedt et al., 2021), with a stronger effect for blue light (Suzuki et al., 2019).

In addition to the finding that blue light enhances the PLR, we found that this effect was specific to occurring above fixation. This result was highly consistent across subjects as observed in the individual correlation maps for the green and blue channels (**Figure 3**, right). While location of fixation was not experimentally manipulated or controlled in this analysis, these

results indicate that fixations toward the horizon may have comprised the majority of fixations. When this well-matched visual pattern of blue on top and terrain on bottom fell onto the retina, it induced robust pupillary responses that accounted for a large degree of variance in the pupil signal by comparison to other areas or features of the environment. One surprising aspect to this finding is that the blue sky itself had lower luminance intensity compared to the terrain below, which was composed of more red and green hues that tend to carry stronger luminance signals (See **Supplementary File 1**). So while the strength of the luminance signal tended to be stronger below fixation, it was the specific temporal pattern of blue luminance above fixation that correlated most with pupil size, indicating the strength of this contextual effect.

Confirming the "Blue Sky Effect"

To confirm this finding and evaluate whether the "blue sky effect" has a more generalized effect on the PLR, we ran a second experiment with a controlled task to assess whether the PLR is indeed stronger, specifically, when blue is seen above fixation. The experimental procedure was careful to match the luminance level across color stimuli (gray, blue, red, and green). The only variables that were manipulated on each trial were the color presented (paired with gray) and its location (on top, bottom, left, or right). Results showed that the PLR was indeed stronger for blue stimuli in comparison to green and red, which replicates results from prior studies (Suzuki et al., 2019). Importantly, the results also showed a specific effect in which the PLR was strongest when blue was presented on top compared to on bottom. There were no other statistically significant effects associated with the relative location of other color stimuli. This result provides additional support to the blue sky hypothesis, suggesting that the PLR is influenced most strongly by blue light, and in particular, when blue light is present above fixation.

An evolutionary and ecological perspective can provide a speculative explanation of the observed blue sky effect on the PLR. A primary function of the pupil is to constrict in response to a sudden bright stimulus and to dilate in response to relatively dark stimuli in order to maintain optimal acuity under a wide range of visual conditions, and perhaps also to protect the photosensitive retina from very strong levels of brightness (Mathôt, 2018). Through lifetimes of experience on earth, a mechanism may have evolved within the circuit that controls the PLR to be particularly sensitive to visual patterns that indicate a person is outdoors in sunny, daytime conditions, and to constrict in anticipation of the subsequent strong levels of brightness (Laeng and Endestad, 2012; Zavagno et al., 2017; Suzuki et al., 2019). While the present study cannot shed light on the purpose or specific mechanisms supporting the observed heightened sensitivity of the pupillary system to blue light from above, future studies may be designed to better ascertain the underlying mechanisms driving this effect. For example, it would be informative to test whether other species show a similar sensitivity to blue light from above or to examine the strength of this effect in special populations, such as those with color blindness or other visual disorders. If instead this effect were merely due to unequal distributions of short wavelength sensitive photoreceptor cells in the retina, then we might expect between subject variability in receptor density to correlate with the strength of the blue light from above effect on the PLR.

Practical Implications

There remain substantial challenges to enabling reliable inference of pupil-linked states in complex or uncontrolled visual environments outside the laboratory due to the fact that multiple neural systems combine to influence the unitary pupillary signal (the sequence of constrictions and dilations over time). Change in pupil size can reflect mental states over a range of temporal scales, from transient processes related to attention (Nieuwenhuis et al., 2011; Preuschoff et al., 2011; Geva et al., 2013; van den Brink et al., 2016) and decision making (Einhauser et al., 2010; Cavanagh et al., 2014; Hoffing et al., 2020) that unfold over a few seconds, to more general arousal and fatigue states (Franklin et al., 2013; Hopstaken et al., 2015; Unsworth and Robison, 2018) that unfold over a longer time period (Aston-Jones and Cohen, 2005). As a result, it is difficult to disentangle whether the pupillary signal at any given point in time reflects the influence of a cognitive or non-cognitive factor (Mathôt, 2018). One potential approach to tackle this challenge is to develop appropriate models to best account for complex luminance signals on pupil size and then examine the residual unaccounted for variance to better estimate cognitive-based effects.

The observed blue light from above effect on pupil size has practical implications for analysis of pupillary data in complex virtual environments and in real-world scenarios. For future, real-world systems that aim to measure environmental light to account for the influence of non-cognitive luminance effects on pupil size, it may be pertinent to ensure that blue light is captured and modeled appropriately relative to other wavelengths to ensure the best prediction of light-induced pupillary fluctuations. It could also be the case that a sensor specifically tuned to blue wavelength light may be sufficient for some applications. There is, however, still much work to fully understand the efficacy of subtractive models that attempt to better characterize cognitive influences on the pupil by subtracting or factoring out estimates of non-cognitive influences. We believe that a major contribution of this work is in rejecting the foveal bias hypothesis for estimating the influence of luminance on pupil size in complex visual environments. It does not appear to be the case that the visual system uses an area around the high density fovea as a sensor to pool luminance information for driving the PLR. The system instead seems more sensitive to global patterns of information and contextual cues, like the blue sky effect and perhaps others, in determining the appropriate pupil response to light in a given environment.

For controlled studies in cognitive pupillometry, this also suggests a prime importance for controlling the level (and visual patterns) associated with blue light in particular. Experimenters should be careful to counterbalance conditions if color stimuli are used with variation along the color spectrum to avoid confounds that might influence interpretation of phasic pupillary responses to cognitive events. Such color spectrum effects should

also be accounted for (or anticipated in the analysis) in these studies as equating luminance alone may not have the desired effect of equating the strength of the pupillary light response to differently colored stimuli. As shown here, the pupil response to a blue stimulus is significantly larger than to a red stimulus with well-matched luminance. As our data show, even the relative location of the blue stimulus (above or below fixation) can impact the strength of the pupil response. Further care should be taken in interpreting pupillary data, especially when examining pupillary responses to naturalistic or more complex stimuli that match the spatial structure of the blue light from above effect (e.g., blue predominantly in the upper part of the image).

Limitations

There are some limitations associated with the current study. The analyses in Experiment 1 focused on examining correlations (i.e., linear associations) between pupil size and RGB luminance over a rather long period of time (up to 15 min of data or 8,000–10,000 contiguous data points). While this approach is well-suited for capturing longer timescale relational trends in the data, this analysis technique does not have sufficient temporal resolution to capture how momentary changes in luminance and/or scene characteristics influence ballistic changes in pupil size at short timescales (e.g., on the order of seconds or milliseconds). Our correlational analysis does show that increases in RGB luminance tend to be associated with reductions in pupil size (and vice versa), as expected, but this approach does not shed light on the precise relationship between sudden changes in complex luminance patterns, whether due to eye movements or image dynamics, and the precise shape of the PLR. It is unclear the extent to which existing parametric models of the PLR to single transient changes in luminance, such as the gamma function (Korn and Bach, 2016), would extend to the rich, dynamic image sequences seen in a virtual environment like the one used in the current experiment. We did incorporate the expected time delay between luminance and pupil size (empirically derived to be about 500 ms) in our correlational analyses, but did not convolve dynamic luminance patterns with a specific parametric impulse response function. It is unclear whether such models would better account for short timescale pupillary fluctuations and impact results of the current study, but poses an interesting question for future research.

Because the task was performed in a virtual environment of our design, which can be described as a mountainous desert wasteland landscape, the set of images seen by subjects on the computer screen tended to have a particular spatial structure and spectral composition. The pattern of results we observed in Experiment 1 are clearly influenced by this structure. For example, the correlation maps show correlation values that vary primarily in the vertical direction, but not the horizontal direction, likely due to the structure of the images themselves. In a horizontal slice of the image near the bottom there will tend to be more red, brown, and grayish colors associated with the terrain and less variability from left to right. A horizontal slice in the upper part of many images would often consist predominantly of blue color associated with the sky, as long as subjects are looking toward the horizon (which was likely a common focal point for subjects as they navigated the virtual environment). If the experiment were run in a different virtual environment or even a real-world environment, it is likely that the structure of the correlation maps would change to reflect the prevailing spatial and spectral structure of those images. While this does not greatly impact the main finding in this paper that supported the contextual effect of the blue sky on pupil size, we would expect differences in the structure of correlation maps depending on properties of the visual environment if future studies use this same approach.

Related to the complexity of the task, there is no doubt that various cognitive processes related to visual search, navigation, covert attention, working memory, etc., were ongoing throughout the task that lasted several minutes. Because, we did not control these factors in the experimental design, as it was designed to be a free and active task, we had no way to account for such factors as covariates in the analysis – a situation that is pervasive in real-world cognitive pupillometry. The perspective of our correlation analysis was to essentially treat these cognitive processes as "noise" in the pupillary signal that would influence pupil size independently (and likely to a lesser extent) with respect to luminance information. We do not believe that cognitive effects could explain the robust negative association we found between pixel luminance and pupil size in this task, nor can they explain the reported blue sky effect in the correlation maps. If anything, the correlations we measured are likely an underestimate of the actual strength of the influence of luminance on pupil size because of the unaccounted for "noise" introduced by continuous cognitive processes.

A subset of our participants were older than 60 years ($n = 4$), and with age there is a tendency for yellowing of the eye's lens and changes in the perception of color, particularly blue light (Gaillard et al., 2000; Michael and Bron, 2011). We re-ran our analyses only for subjects with ages less than 60 years ($n = 30$) and found that the pattern of results and level of statistical significance was not greatly impacted. In fact, all four of the subjects over 60 years showed a blue sky effect (significantly stronger negative correlation for blue pixels above fixation versus below). A recent study by Rukmini et al. (2017) showed that while the strength of the pupillary light response is reduced overall with aging, the amount of reduction was similar for red (631 nm) and blue (469 nm) wavelength light. They concluded that yellowing of the lens does not selectively reduce melanopsin-dependent light responses as reflected by the pupillary light response for people over 60 years old. Our data are consistent with this finding.

Lastly, in Experiment 2, we were not able to perfectly equate luminance across colored stimuli (**Figure 6C**). Based on the measurements, we obtained with the spectroradiometer in the lab, we were only able to match luminance in a range from 16.6 to 19.4 cd/m^2, which is a very narrow range compared to the full spectrum (which ranges from 0–240 cd/m^2), but the possibility remains that the response to blue stimuli was larger than red in Experiment 2 because of the luminance difference of 2.8 cd/m^2. The dose-response curve is not well mapped out for PLR amplitude and incremental changes in

luminance for different wavelengths of light, so we believe the possibility remains that the difference between red and blue could be due to luminance *per se*, and not a color specific phenomenon in Experiment 2. We will note that this potential confound does not affect our interpretation of the orientation-specific blue sky effect because we compared PLRs to stimuli within each color that were equiluminant and only varied by the location of the color.

CONCLUSION

In this study, we examined correlations between pupil size and dynamic image statistics in the context of a free visual search and navigation task in a 3D virtual environment. We found that blue and green pixel intensities had a disproportionately large impact on pupil size in comparison with the red color channel. Furthermore, we found that visual scenes in which blue was predominantly overhead had the strongest influence on pupil size, which led us to hypothesize a "blue sky effect." We conducted a follow-up controlled laboratory experiment and found evidence consistent with our hypothesis, showing a specific sensitivity of the PLR to blue light when it is located above fixation. From an ecological perspective, we speculate that the heightened sensitivity of the pupillary system to this visual pattern may be a useful adaptive response due to the persistent association between sunlight, large increases in brightness, and the blue sky in our daily lives. From a practical standpoint in terms of pupillometry research, the findings of this report suggest that equating luminance alone may be insufficient to account for luminance effects on pupil size if multi-colored stimuli and/or naturalistic images are used in psychological research. More research is necessary to fully understand how best to account for the influence of light on pupil size for studies or applications in complex visual environments outside the laboratory.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available upon request via email by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ARL Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ST conducted the analysis for both experiments. SG and JT designed Experiment 1. ST and RH designed Experiment 2 and wrote the first draft of the manuscript. AM and AR collected data for Experiment 2. AM, AR, JT, and SG contributed to manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

This research was sponsored by the Army Research Laboratory and was accomplished under Contract Number 1782 W911NF-10-D-0002.

ACKNOWLEDGMENTS

We would like to thank Ashley Oiknine and Bianca Delangin for their work with data collection and subject recruitment, and Min Wei for work in developing the virtual environment and task. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the United States Government. The United States Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.748539/full#supplementary-material

REFERENCES

Aston-Jones, G., and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* 28, 403–450. doi: 10.1146/annurev.neuro.28.061604.135709

Barrionuevo, P. A., and Cao, D. (2016). Luminance and chromatic signals interact differently with melanopsin activation to control the pupil light response. *J. Vis.* 16:29. doi: 10.1167/16.11.29

Barrionuevo, P. A., Nicandro, N., McAnany, J. J., Zele, A. J., Gamlin, P., and Cao, D. (2014). Assessing rod, cone, and melanopsin contributions to human pupil flicker responses. *Invest. Ophthalmol. Vis. Sci.* 55, 719–727. doi: 10.1167/iovs.13-13252

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi: 10.2307/2346101

Bonmati-Carrion, M. A., Hild, K., Isherwood, C., Sweeney, S. J., Revell, V. L., Skene, D. J., et al. (2016). Relationship between human pupillary light reflex and circadian system status. *PLoS One.* 11, e0162476.

Bonmati-Carrion, M. A., Hild, K., Isherwood, C. M., Sweeney, S. J., Revell, V. L., Madrid, J. A., et al. (2018). Effect of single and combined monochromatic light on the human pupillary light response. *Front. Neurol.* 9:1019. doi: 10.3389/fneur.2018.01019

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897X00357

Castellotti, S., Conti, M., Feitosa-Santana, C., and Del Viva, M. M. (2020). Pupillary response to representations of light in paintings. *J. Vis.* 20:14. doi: 10.1167/jov.20.10.14

Cavanagh, J. F., Wiecki, T. V., Kochar, A., and Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *J. Exp. Psychol. Gen.* 143, 1476–1488. doi: 10.1037/a0035813

Crawford, B. H., and Parsons, J. H. (1936). The dependence of pupil size upon external light stimulus under static and variable conditions. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 121, 376–395. doi: 10.1098/rspb.1936.0072

Curcio, C. A., Allen, K. A., Sloan, K. R., Lerea, C. L., Hurley, J. B., Klock, I. B., et al. (1991). Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *J. Comp. Neurol.* 312, 610–624. doi: 10.1002/cne.903120411

Czeisler, C. A., Shanahan, T. L., Klerman, E. B., Martens, H., Brotman, D. J., Emens, J. S., et al. (1995). Suppression of melatonin secretion in some blind patients by exposure to bright light. *N. Engl. J. Med.* 332, 6–11. doi: 10.1056/NEJM199501053320102

Delorme, A., Mullen, T., Kothe, C., Acar, Z. A., Bigdely-Shamlo, N., Vankov, A., et al. (2011). EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing. *Comput. Intell. Neurosci.* 2011:130714. doi: 10.1155/2011/130714

Denison, R. N., Parker, J. A., and Carrasco, M. (2020). Modeling pupil responses to rapid sequential events. *Behav. Res. Methods* 52, 1–17. doi: 10.3758/s13428-020-01368-6

Do, M. T. H., and Yau, K.-W. (2010). Intrinsically photosensitive retinal ganglion cells. *Physiol. Rev.* 90, 1547–1581. doi: 10.1152/physrev.00013.2010

Dunn, O. J., and Clark, V. (1969). Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc.* 64, 366–377. doi: 10.1080/01621459.1969.10500981

Einhauser, W., Koch, C., and Carter, O. L. (2010). Pupil dilation betrays the timing of decisions. *Front. Hum. Neurosci.* 4:18. doi: 10.3389/fnhum.2010.00018

Ellis, C. J. (1981). The pupillary light reflex in normal subjects. *Br. J. Ophthalmol.* 65, 754–759. doi: 10.1136/bjo.65.11.754

Enders, L. R., Smith, R. J., Gordon, S. M., Ries, A. J., and Touryan, J. (2021). Gaze behavior during navigation and visual search of an open-world virtual environment. *Front. Psychol.* 12:681042. doi: 10.3389/fpsyg.2021.681042

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., and Schooler, J. W. (2013). Window to the wandering mind: pupillometry of spontaneous thought while reading. *Q. J. Exp. Psychol.* 66, 2289–2294. doi: 10.1080/17470218.2013.858170

Frazor, R. A., and Geisler, W. S. (2006). Local luminance and contrast in natural images. *Vis. Res.* 46, 1585–1598. doi: 10.1016/j.visres.2005.06.038

Gaillard, E. R., Zheng, L., Merriam, J. C., and Dillon, J. (2000). Age-related changes in the absorption characteristics of the primate lens. *Invest. Ophthalmol. Vis. Sci.* 41, 1454–1459

Geva, R., Zivan, M., Warsha, A., and Olchik, D. (2013). Alerting, orienting or executive attention networks: differential patters of pupil dilations. *Front. Behav. Neurosci.* 7:145. doi: 10.3389/fnbeh.2013.00145

Gosselin, F., and Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vis. Res.* 41, 2261–2271. doi: 10.1016/S0042-6989(01)00097-9

Hoffing, R. A. C., Lauharatanahirun, N., Forster, D. E., Garcia, J. O., Vettel, J. M., and Thurman, S. M. (2020). Dissociable mappings of tonic and phasic pupillary features onto cognitive processes involved in mental arithmetic. *PLoS One* 15:e0230517. doi: 10.1371/journal.pone.0230517

Hopstaken, J. F., van der Linden, D., Bakker, A. B., and Kompier, M. A. (2015). The window of my eyes: task disengagement and mental fatigue covary with pupil dynamics. *Biol. Psychol.* 110, 100–106. doi: 10.1016/j.biopsycho.2015.06.013

Hu, X., Hisakata, R., and Kaneko, H. (2020). Effects of stimulus size, eccentricity, luminance, and attention on pupillary light response examined by concentric stimulus. *Vis. Res.* 170, 35–45. doi: 10.1016/j.visres.2020.03.008

Joshi, S., and Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends Cogn. Sci.* 24, 466–480. doi: 10.1016/j.tics.2020.03.005

Kelbsch, C., Stingl, K., Kempf, M., Strasser, T., Jung, R., Kuehlewein, L., et al. (2019). Objective measurement of local rod and cone function using gaze-controlled chromatic pupil campimetry in healthy subjects. *Transl. Vis. Sci. Technol.* 8:19. doi: 10.1167/tvst.8.6.19

Korn, C. W., and Bach, D. R. (2016). A solid frame for the window on cognition: modeling event-related pupil responses. *J. Vis.* 16:28. doi: 10.1167/16.3.28

Kothe, C. (2014). Lab streaming layer (LSL). Swartz Center for Computational Neuroscience. Available at: https://github.com/sccn/labstreaminglayer (Accessed 11 June, 2021).

Kothe, C. A., and Makeig, S. (2013). BCILAB: a platform for brain–computer interface development. *J. Neural Eng.* 10:056014. doi: 10.1088/1741-2560/10/5/056014

Laeng, B., and Endestad, T. (2012). Bright illusions reduce the eye's pupil. *Proc. Natl. Acad. Sci. U. S. A.* 109, 2162–2167. doi: 10.1073/pnas.1118298109

Lee, S., Muto, N., Shimomura, Y., and Katsuura, T. (2017). Human pupillary light reflex during successive irradiation with 1-ms blue-and green-pulsed light. *J. Physiol. Anthropol.* 36, 1–7. doi: 10.1186/s40101-017-0153-7

Legras, R., Gaudric, A., and Woog, K. (2018). Distribution of cone density, spacing and arrangement in adult healthy retinas with adaptive optics flood illumination. *PLoS One* 13:e0191141. doi: 10.1371/journal.pone.0191141

Lenhard, W., and Lenhard, A. (2014). Hypothesis Tests for Comparing Correlations. (Germany: Psychometrica available: https://www.psychometrica.de/correlation.html. Bibergau. doi: 10.13140/RG.2.1.2954.1367

Lockley, S. W., Skene, D. J., Arendt, J., Tabandeh, H., Bird, A. C., and Defrance, R. (1997). Relationship between melatonin rhythms and visual loss in the blind. *J. Clin. Endocrinol. Metab.* 82, 3763–3770. doi: 10.1210/jcem.82.11.4355

Lucas, R. J., Douglas, R. H., and Foster, R. G. (2001). Characterization of an ocular photopigment capable of driving pupillary constriction in mice. *Nat. Neurosci.* 4, 621–626. doi: 10.1038/88443

Lucas, R. J., Hattar, S., Takao, M., Berson, D. M., Foster, R. G., and Yau, K.-W. (2003). Diminished pupillary light reflex at high irradiances in melanopsin-knockout mice. *Science* 299, 245–247. doi: 10.1126/science.1077293

Mathôt, S. (2013). A simple way to reconstruct pupil size during eye blinks. Retrieved from https://doi.org/10.6084/m9.figshare.688001

Mathôt, S. (2018). Pupillometry: psychology, physiology, and function. *J. Cogn.* 1:16. doi: 10.5334/joc.18

McDougal, D. H., and Gamlin, P. D. (2010). The influence of intrinsically-photosensitive retinal ganglion cells on the spectral sensitivity and response dynamics of the human pupillary light reflex. *Vis. Res.* 50, 72–87. doi: 10.1016/j.visres.2009.10.012

Michael, R., and Bron, A. J. (2011). The ageing lens and cataract: a model of normal and pathological ageing. *Philos. Trans. R. Soc. B Biol. Sci.* 366, 1278–1292. doi: 10.1098/rstb.2010.0300

Naber, M., and Nakayama, K. (2013). Pupil responses to high-level image content. *J. Vis.* 13:7. doi: 10.1167/13.6.7

Neitz, J., and Neitz, M. (2011). The genetics of normal and defective color vision. *Vis. Res.* 51, 633–651. doi: 10.1016/j.visres.2010.12.002

Nieuwenhuis, S., De Geus, E. J., and Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology* 48, 162–175. doi: 10.1111/j.1469-8986.2010.01057.x

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442. doi: 10.1163/156856897X00366

Preuschoff, K., Hart, B. M., and Einhauser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Front. Neurosci.* 5:115. doi: 10.3389/fnins.2011.00115

Rukmini, A. V., Milea, D., Aung, T., and Gooley, J. J. (2017). Pupillary responses to short-wavelength light are preserved in aging. *Sci. Rep.* 7:43832. doi: 10.1038/srep43832

Sirois, S., and Brisson, J. (2014). Pupillometry. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 679–692. doi: 10.1002/wcs.1323

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 245–251. doi: 10.1037/0033-2909.87.2.245

Stockman, A., MacLeod, D. I., and Johnson, N. E. (1993). Spectral sensitivities of the human cones. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 10, 2491–2521. doi: 10.1364/JOSAA.10.002491

Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: a review. *J. Vis.* 11:13. doi: 10.1167/11.5.13

Sulutvedt, U., Zavagno, D., Lubell, J., Leknes, S., de Rodez Benavent, S. A., and Laeng, B. (2021). Brightness perception changes related to pupil size. *Vis. Res.* 178, 41–47. doi: 10.1016/j.visres.2020.09.004

Suzuki, Y., Minami, T., Laeng, B., and Nakauchi, S. (2019). Colorful glares: effects of colors on brightness illusions measured with pupillometry. *Acta Psychol.* 198:102882. doi: 10.1016/j.actpsy.2019.102882

Thurman, S. M., Giese, M. A., and Grossman, E. D. (2010). Perceptual and computational analysis of critical features for biological motion. *J. Vis.* 10:15. doi: 10.1167/10.12.15

Thurman, S. M., and Lu, H. (2013). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychol. Sci.* 24, 1133–1141. doi: 10.1177/0956797612467212

Unsworth, N., and Robison, M. K. (2018). Tracking arousal state and mind wandering with pupillometry. *Cogn. Affect. Behav. Neurosci.* 18, 638–664. doi: 10.3758/s13415-018-0594-4

van den Brink, R. L., Murphy, P. R., and Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLoS One* 11:e0165274. doi: 10.1371/journal.pone.0165274

Williams, E. J. (1959). The comparison of regression variables. *J. R. Stat. Soc. Ser. B Methodol.* 21, 396–399. doi: 10.1111/j.2517-6161.1959.tb00346.x

Wood, M. (2012). Lightness-the helmholtz-kohlrausch effect. Out of the Wood 20–22.

Zavagno, D., Tommasi, L., and Laeng, B. (2017). The eye pupil's response to static and dynamic illusions of luminosity and darkness. *Iperception* 8:2041669517717754. doi: 10.1177/2041669517717754

Conflict of Interest: SG is employed by DCS Corporation (United States).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

METHODS
published: 31 December 2021
doi: 10.3389/fpsyg.2021.650693

A Case for Studying Naturalistic Eye and Head Movements in Virtual Environments

Chloe Callahan-Flintoft[1], Christian Barentine[2], Jonathan Touryan[1] and Anthony J. Ries[1,2]*

[1]Humans in Complex System Directorate, United States Army Research Laboratory, Adelphi, MD, United States, [2]Warfighter Effectiveness Research Center, United States Air Force Academy, Colorado Springs, CO, United States

Using head mounted displays (HMDs) in conjunction with virtual reality (VR), vision researchers are able to capture more naturalistic vision in an experimentally controlled setting. Namely, eye movements can be accurately tracked as they occur in concert with head movements as subjects navigate virtual environments. A benefit of this approach is that, unlike other mobile eye tracking (ET) set-ups in unconstrained settings, the experimenter has precise control over the location and timing of stimulus presentation, making it easier to compare findings between HMD studies and those that use monitor displays, which account for the bulk of previous work in eye movement research and vision sciences more generally. Here, a visual discrimination paradigm is presented as a proof of concept to demonstrate the applicability of collecting eye and head tracking data from an HMD in VR for vision research. The current work's contribution is 3-fold: firstly, results demonstrating both the strengths and the weaknesses of recording and classifying eye and head tracking data in VR, secondly, a highly flexible graphical user interface (GUI) used to generate the current experiment, is offered to lower the software development start-up cost of future researchers transitioning to a VR space, and finally, the dataset analyzed here of behavioral, eye and head tracking data synchronized with environmental variables from a task specifically designed to elicit a variety of eye and head movements could be an asset in testing future eye movement classification algorithms.

Keywords: head mounted display, eye tracking, eye movement analysis, virtual reality, smooth pursuit

OPEN ACCESS

Edited by:
Andrey R. Nikolaev, Lund University, Sweden

Reviewed by:
Vsevolod Peysakhovich, Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), France
Ulrik Günther, Helmholtz Association of German Research Centers (HZ), Germany

****Correspondence:***
Chloe Callahan-Flintoft
ccallahanflintoft@gmail.com

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *08 January 2021*
Accepted: *10 November 2021*
Published: *31 December 2021*

Citation:
Callahan-Flintoft C, Barentine C, Touryan J and Ries AJ (2021) A Case for Studying Naturalistic Eye and Head Movements in Virtual Environments. Front. Psychol. 12:650693. doi: 10.3389/fpsyg.2021.650693

INTRODUCTION

Understanding how the visual system operates in the natural environment is a fundamental goal of cognitive psychology and has consequences for a variety of other research fields such as human factors and advertising. The natural environment offers a complex and uncontrolled input of visual information, making it is difficult to isolate variables of interest and determine their effect on behavior. Alternatively, constrained laboratory experimentation offers precise control, while potentially limiting the generalizability to less confined environments. To this end, vision researchers have begun to strike a balance between the lab and real world by running experiments in virtual reality (VR) using head mounted displays (HMDs). Experimentation in VR enables research paradigms that allow for more naturalistic behavior in subjects, while still providing experimental control over stimulus presentation (Clay et al., 2019).

For purposes of clarity, we refer to the three-dimensional virtual environment, the digital X, Y, Z space in which one can present stimuli, *via* game engines such as Unity or Unreal, as VR (Watson et al., 2019). HMD refers specifically to the video display worn on the head, where subjects are immersed in a 360° virtual environment. Using VR in combination with HMDs allows experimenters precise control over stimulus timing and location, while offering subjects superior (when compared to traditional computer monitor setups) depth perception, a wider field of view, and the ability to move the eyes and head as they would in the real world (see Discussion for a more nuanced discussion of the limitations of VR and HMDs).

Integrating eye tracking (ET) with HMDs has further extended the research potential for this technology (Jangraw et al., 2014). Eye tracking has been an essential component in understanding how the visual system acquires information from a scene to build our internal perception. Measuring where the eyes foveate in a scene has been demonstrated in VR with HMD systems (Clay et al., 2019) and has given insight into how scene gist influences eye fixations during search (Boettcher et al., 2018) as well as how bottom-up and top-down influences guide the deployment of attention (Anderson et al., 2015; Harada and Ohyama, 2019). How the eyes move around a scene also provides important insight into cognitive processes (Williams and Castelhano, 2019) as well as clinical applications (Baloh et al., 1975; Terao et al., 2017; Ward and Kapoula, 2020). However, the vast majority of this research has been performed using a camera-based eye tracker and a two-dimensional monitor, which restricts the space stimuli are presented in and the subsequent behavior they induce. Newer technologies such eye tracking enabled HMDs, in addition to eye tracking glasses (ETGs), provide access to similar data but with the added benefit of tracking gaze in a 360° environment.

A growing effort has been made to study vision using more naturalistic scenes (Henderson et al., 2007; Dorr et al., 2010; Wolfe et al., 2011; O'Connell and Chun, 2018). However, equally important to exploring vision in the context of natural input (i.e., real-world scenes), is to explore vision in tandem with natural movement. Both HMD with VR and ETGs offer the freedom to move the head and torso when viewing the environment. ETGs have the added benefit of also allowing the subject to walk around the environment unrestricted, whereas subjects are typically more limited in HMDs, having to rely on unnatural modes of transport such as teleportation to avoid collisions with physical objects and to maintain a position within the headset-tracking volume. However, HMDs do allow more natural movement on smaller scales (e.g., room-size) and ETGs do not control for stimulus presentation that can be variable and unpredictable in real environments and may be less viable in situations such as training, where *in situ* exposure could be dangerous and/or costly (e.g., a simulated battlefield). In either circumstance, the ability to quantify more complex and dynamic eye movement patterns observed is limited as the majority of classification algorithms were developed with static 2D stimuli, and do not generalized to naturalistic contexts (Agtzidis et al., 2020).

The current work uses a visual discrimination task with unrestricted eye and head movement. Elicited patterns of activity are then classified based on the thresholds of I-S^{5}T, which thresholds eye, head, and gaze (eye + head) speed (Agtzidis et al., 2019). The original thresholding system was simplified to classify the following types of eye movements: saccades (a high-speed ballistic eye movement), fixation (a period of low to no eye speed), smooth pursuit (a period where the eyes are moving to foveate a moving stimulus), VOR (a period of low to no gaze speed but the eyes are moving in the head to compensate for head motion), and head pursuit (a period of low to no eye speed but gaze is moving, driven by head motion in order to foveate a moving stimuli). A secondary form of smooth pursuit was also classified as smooth pursuit with compensatory VOR (a period where gaze is moving to foveate a moving target and the eyes are moving relative to the head to compensate for head motion). This classification is tested during temporal epochs when smooth pursuit eye movements are likely (i.e., when subjects must track a moving stimulus) and compared to other epochs when smooth pursuit is unlikely (i.e., when the eyes must foveate a static object). As opposed to previous work that has used eye tracking with HMDs, presenting more complex or naturalistic scenes (e.g., 360° videos; Rai et al., 2017; Haskins et al., 2020; Kim et al., 2020), here, stimulus presentation is strictly controlled while viewing behavior (i.e., the movement of the eyes and head) is not. The use of simplified stimuli, similar to those used in previous, 2D display paradigms, allows for an easier comparison to previous results in order to explore how head movements may interact with the execution of eye movements or underlying cognitive processes. The paradigm presented here is generated using a graphical user interface (GUI) specifically designed to allow future researchers to adapt stimulus parameters such as eccentricity and motion speed, in the continued effort to understand how well-studied eye movement phenomena may or may not change when subjects' viewing is less restricted. The strict control of stimulus presentation is meant to elicit predictable eye movements such as saccades and smooth pursuit in the presence of head motion. This, coupled with the ground truth knowledge of the location of stimuli relative to the viewer's gaze direction, makes this dataset uniquely beneficial to the development of more automated eye movement classification algorithms.

MATERIALS AND METHODS

Ethics Statement

This experiment was approved by the Institutional Review Board at the United States Air Force Academy (USAFA) and United States Army Research Laboratory (ARL) under Project Number ARL 19–122. All procedures were in accordance with the Declaration of Helsinki.

Subjects

Twenty-four subjects (United States Air Force Academy cadets; nine female, average age 19.3 years) were tested and received course credit for their participation. Subjects were recruited through Sona Systems and provided written informed consent prior to experimentation. All subjects had normal or corrected to normal vision.

Apparatus

Experimental procedures were designed using the Unity gaming engine.[1] Stimuli were presented to an HTC Vive VR headset (1,080 × 1,200 pixels per eye, 90 Hz refresh rate, 110° field of view) with integrated eye tracking from Tobii Technologies (120 Hz sampling rate, Tobii Pro SDK) using a Corsair One PC (Windows 10, Intel Core i9 CPU @ 3.6GHz, 64-bit, Nvidia GeForce RTX 2080Ti, 32 GB RAM) and two external lighthouses used for tracking head position. Subjects were given instructions and practiced to correctly position the VR headset prior to experimentation. Subjects were comfortably seated in a fixed position chair. The Tobii Vive was used here as it is a fully integrated system with an estimated accuracy of 0.5°. Other systems such as the Pupil Labs eye tracker can be added to HMD systems and offer higher tracking frequency (200 Hz); however, there is slightly poorer tracking accuracy tracking (1.0°).

Lab Streaming Layer (LSL; available here: https://github.com/labstreaminglayer/LSL4Unity) was used to synchronize eye and head tracking data with button responses and stimulus presentation. LSL is a network-based recording software designed to integrate multiple data streams with sub millisecond precision (Kothe, 2014).

Calibration

The standard five-point calibration contained in the Tobii Pro SDK was implemented before each block of trials. Calibration points were sequentially presented, one each at the four corners of an imaginary square and the middle point centered on the subjects forward gaze position (in Unity meters: corner points +/−0.3x, +/−0.15y, 1.2z, middle point 0, 0, 1.2). Subjects fixated the center of each point, which started at 0.1 m (4.77 degrees of visual angle, dva) in diameter, and shrunk down over the course of fixation until it became invisible, indicating a successfully registered calibration point. Each of the four calibration points was positioned 15.62 dva relative to the middle point. All points were presented in an orthogonal plane at a fixed distance of 1.2 m. If fixation was interrupted during calibration or the calibration point did not disappear, calibration was restarted. While, we did not record the number of calibration attempts or subsequent validation of the calibration, trials did not start until a successful calibration (i.e., all five points were fixated and registered by the software as completed) was accomplished. No subjects were removed due to poor calibration.

Units of Measurement in Virtual Environments

Unity objects (stimuli) are defined in world coordinate using notional or approximate meters. However, a more precise and useful metric for vision scientists is dva. As such, both are reported in this paper. It is important to note that the degrees of visual angle are approximate. The screen inside the headset does not fully cover the natural human field of view, leading to a small binocular effect. The fields of view of the virtual cameras are manipulated to counter this effect to make using the headset more comfortable, at the cost of some slight size distortion.

Graphical User Interface for Paradigm Creation

To make this paradigm adaptable for future research, the GUI was included in the software development (**Figure 1A**). Researchers using the supplied code can leverage a GUI to change multiple parameters related to target size, target position, movement speed, quantity, randomization, trial numbers, and temporal contingencies.[2] For example, researchers can set the size and rotation of targets. Likewise, researchers can change the perceived motion from an observer moving through a space of static disks to the disks moving past a static observer by changing whether the background moves with the participant or with the disks. Additionally, the GUI provides a number of status checks such as indicators that the eye tracker and hand controllers are connected. The intention behind the creation of this GUI was to lower the bar of entry for future researchers and provide a highly flexible generator of visual search or discrimination tasks. The parameters, cited below, were those used in the data reported here. Additionally, all parameters available in the GUI as well as a list of recorded data can be found in the **Supplementary Materials**.

Visual Parameters and Trial Structure

In the virtual environment, a directional light (RGB: 1, 0.95, and 0.84) was used, rotated 50° in the x-axis and −30° in the y-axis in the Unity coordinate system. This had the effect of lighting the scene from over the subject's right shoulder, directed toward their left foot. This ensured all target surfaces were lit. Furthermore, the light created shadows from surrounding objects to provide a sense of depth and create a realistic perception of motion in Dynamic trials (see details below).

Subjects performed 288 trials of a cued, two-alternative forced choice, target discrimination task. On each trial, subjects were instructed to foveate a recentering cross (0.18 m; 0.94 dva) placed 11 m in front of them. The recentering cross was surrounded by an array of white disks (RGB: 1,1,1), each 1 m in diameter evenly spaced on an imaginary circle (**Figure 1B**). Each of the surrounding white disks had the letter "O" at its center (RGB: 0,0,0). On each trial, one of the white disks was cued by turning yellow (RGB: 1, 0.92, and 0.016). Upon the appearance of the cue, subjects were instructed to make a

[1]U. Technologies. Unity – game engine. http://unity3d.com/

[2]https://osf.io/p8g94/

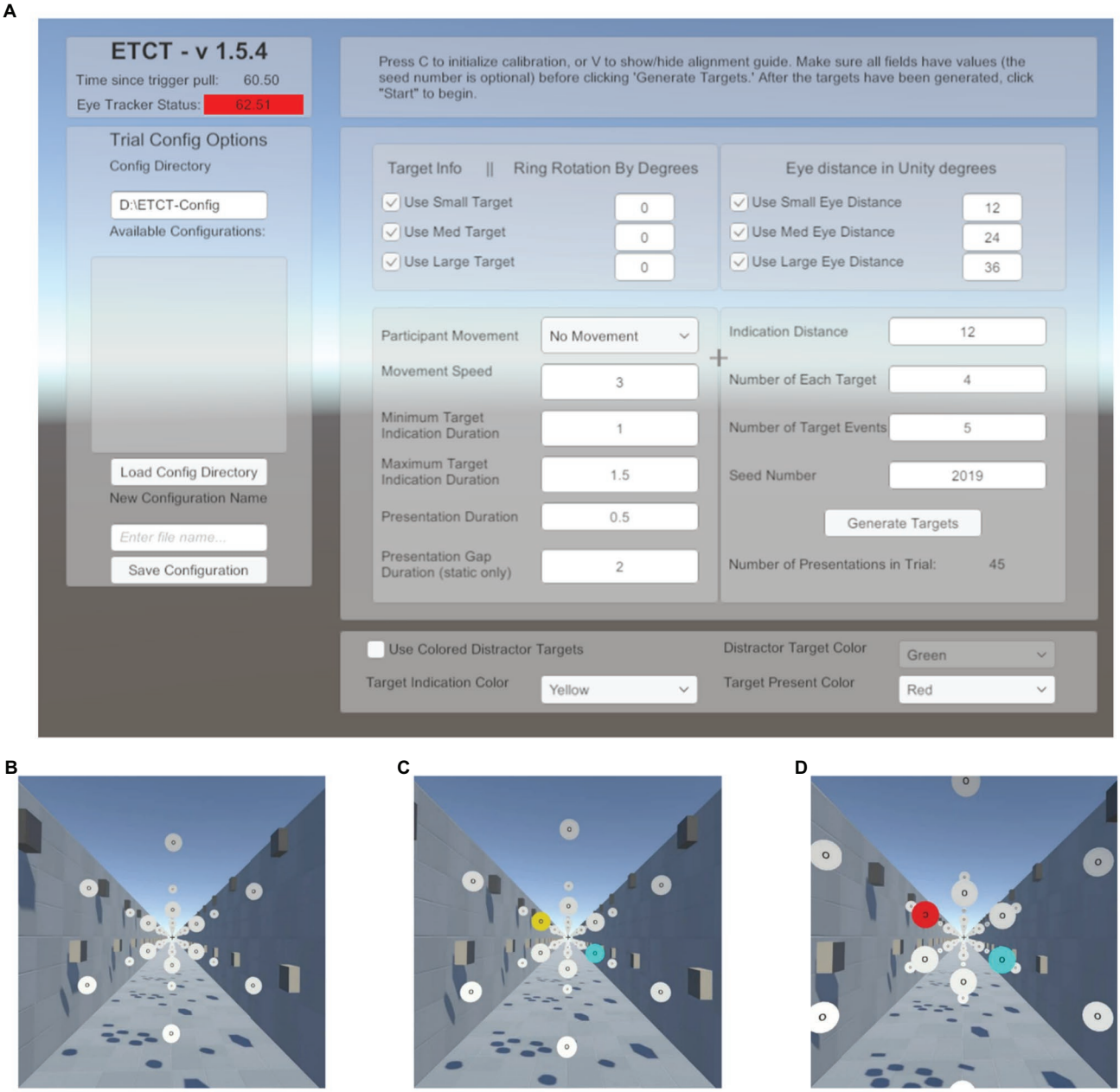


FIGURE 1 | Paradigm schematic. **(A)** Graphical user interface (GUI) presented at the start of the experiment. Here, the researcher can input a number of parameters to alter stimulus presentation. This list includes target parameters such as size as well as the motion (or lack thereof) of the participant through the environment. A list of available parameters and the values set in the current experiment is included in the **Supplementary Materials**. **(B)** Start of a trial. Subjects foveate the recentering cross ahead of them. **(C)** Target indication phase. Subjects are told to saccade to the yellow disk in preparation for the target and ignore the cyan disk. **(D)** Target presentation. Subjects must maintain fixation on the yellow disk until the target is presented, at which time the yellow disk turns red and the "O" at the center of the disk is replaced with a "C" facing either the left or the right. Subjects report the direction of the "C" using the virtual reality (VR) controls and return their gaze to the recentering cross to await the next trial.

saccade to the "O" in the center of the yellow disk as quickly as possible and foveate the center until a subsequent target was presented. The presentation of the cue was not gaze-contingent meaning it would occur even if the subject's eyes were not on the recentering cross. Simultaneously with the yellow cue presentation, a counterpart disk, at a diametrically opposite point along the ring relative to the cued disk, turned cyan (RGB: 0,1,1; **Figure 1C**). The cyan distractor disk was included for the purposes of piloting for a follow-up study using EEG in order to control sensory input between visual fields and prevent reflexive saccades. Our analyses focus only on the target disk and thus the counterpart cyan disk is not discussed further in this paper. After 600–1,600 ms the yellow disk turned red (RGB: 1,0,0) and simultaneously the "O" label

was replaced with the target, a "C," faced either to the left or right (**Figure 1D**). The interval between cue (yellow disk) and target (red with "C") was used to allow subjects enough time to locate and saccade to the disk as well as provide variable of tracking the disk, relative to target onset. Subjects were told to use the VR controllers, one in each hand, to report if the "C" was open to the left or the right, an equally probable occurrence that required responses from the left and right controllers, respectively. The "C" was present for 1,000 ms. Subjects were instructed to return their gaze to the recentering cross as soon as a response was given. There was an average total of 6.4 s between the start of one trial (the onset of the cue) and the next. Responses were counted as valid if they occurred between the presentation of the target and the onset of the cue in the next trial. Only first responses were analyzed.

Subjects experienced both Static and Dynamic trials. In Dynamic trials, subjects "moved" through the environment at 5 m/s. This movement was strictly in the virtual environment as subjects remained stationary in their chair throughout the experiment. The perception of motion was induced by controlling the lighting/shadows in the environment and moving the point of view camera through space. Dynamic trials were included in order to elicit smooth pursuit eye movements. The speed was chosen as a balance between wanting the eyes to move quickly enough to elicit smooth pursuit but not so fast that there was not ample tracking time before a cued disk passed out of view of the participant.

Disks were always cued (turned yellow) 32 m from the subject in the Dynamic trials. In the Static condition trials, the disks were stationary and were cued either 13 or 32 m from subjects. This translated to the disk being approximately 4.4 dva when cued at its closest (13 m) location and 1.8 dva at the farthest (32 m), relative to the subject. Consequently, the "O" (and subsequent "C") on the disks were then 1.89 dva at 13 m from participants and 0.77 dva at 32 ms from participants. The two cueing distances were used to make Static trials more comparable to Dynamic where the cue traveled closer to the subject throughout the trial. Disks were cued in the periphery (20 dva from the recentering cross) and the parafovea (6 dva from the recentering cross) in both Dynamic and Static trials. Target disks and cue durations were randomly selected using a random seed generator. All subjects performed the task using the same seed. That is, each subject experienced the same random order. Subjects performed two blocks of 48 trials in the Static condition, where the cued disk was 13 m from the subject, the Static condition where the cued disk was 32 m from the subject, and the Dynamic condition. The block order was counterbalanced. Calibration was performed prior to each block. Data from each participant can be found online at https://osf.io/p8g94/.

Simulator Sickness Questionnaire

A concern in VR with HMD experimentation is inducing simulator sickness in subjects due to the discrepancy between task-induced motion in the virtual environments and the lack of motion in the real environment. Simulation sickness was measured using the Simulator Sickness Questionnaire (SSQ; Kennedy et al., 1993) before and after the experiment. While the SSQ was designed to measure simulator sickness in flight simulators many researchers have adopted its use in VR environments (see Saredakis et al., 2020, for a review). Subjects rated 16 symptoms on a four-point scale (0–3), which were factored into three categories (Oculomotor, Disorientation, and Nausea) and computed into a Total score.

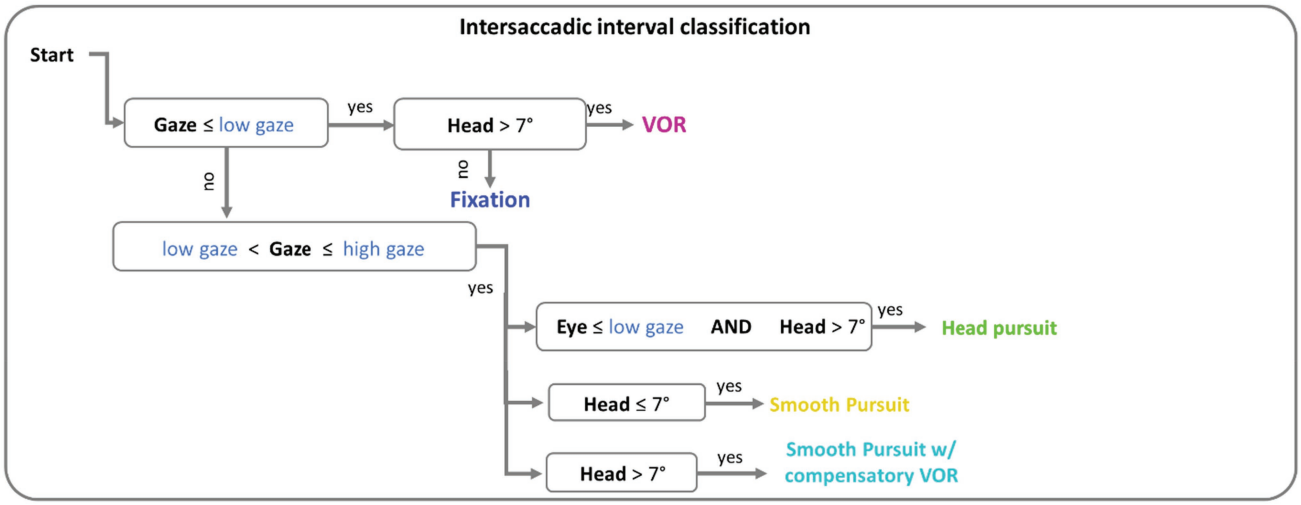
Eye Movement Classification and Validation

Tobii interpolates eye position coordinates for dropped samples (e.g., blinks or missing eye image) but does not interpolate pupil recordings. Therefore, valid eye position values were defined as timepoints, which had corresponding valid pupil samples. Only valid eye position samples were included in analysis. Blinks were not explicitly defined other than as dropped or invalid samples. All invalid epochs, as well as 40 ms before and after the invalid epoch, were considered noise (i.e., invalid) and excluded from classification.

Here, the term eye speed refers to the angular velocity of the eye, relative to the head. Gaze speed refers to angular velocity of foveation relative to the world (the combined eye and head speed). Before classifications of eye movements were made, a five-sample (40 ms) median filter was applied to smooth both eye and head speed data (Engbert and Kiegl, 2003; Engbert and Mergenthaler, 2006; Dimigen et al., 2011). Eye movements were classified by applying a dynamic threshold to gaze and eye speed that is scaled by the current head speed: $\text{threshold}_{\text{scaled}} = (1 + v_{\text{head}}/60)^{*}\text{threshold}$, where v_{head} is the velocity of the head at a given time point (see Agtzidis et al., 2019, for details). Saccade detection was performed first. The label "saccade" was applied to all time points in windows of 20 ms or longer, where eye speed exceeded the scaled velocity threshold [for saccades this would be $(1 + v_{\text{head}}/60)^{*}\theta_{\text{Saccade}}$, where θ_{Saccade} is the saccade threshold when the head is stationary, 35 deg/s; see **Figure 2**]. For analysis, only saccades over 3° in amplitude with a peak velocity under 1,000 deg./s were included. This velocity cutoff was based on previous research (Holmqvist et al., 2011; Ries et al., 2018) to exclude improbable eye movements, and 3° was used to exclude small eye movements around the recentering cross.

Intersaccadic intervals were classified in 100 ms epochs based on a set of thresholds (see **Figure 2**) for the gaze speed and head speed. The implementation of thresholding here has been outlined in the flow chart of **Figure 2**. If gaze speed was below the scaled low gaze threshold [that would be calculated as $(1 + v_{\text{head}}/60)^{*}\theta_{\text{lowgaze}}$, where θ_{lowgaze} is the lower bound gaze threshold when the head is stationary, 10 deg./s] then the window was assigned a label of "VOR" or "fixation" depending on if the head above threshold (7 deg./s). If gaze was moving (above scaled low threshold), the epoch was classified as a "head pursuit" if the eye speed was below threshold, "smooth pursuit" if the head speed was below threshold, or "smooth pursuit with compensatory VOR" if both head and eye speed were above threshold.

The classification algorithm was compared against ray-casts of each gaze sample. Ray-casting is when an imaginary ray



Name	Used for	Threshold
θ saccade	Saccades	35°/sec
θ low gaze	Fixation, smooth pursuit, VOR, head pursuit	10°/sec
θ high gaze	Fixation, smooth pursuit, VOR, head pursuit	65°/sec
θ low head	VOR, head pursuit	7°/sec
θ high head	Scaling θ for low/high gaze	60°/sec

FIGURE 2 | Flow chart of the threshold classification. Intersaccadic intervals were divided into 100 ms windows for classification. Threshold in blue font are thresholds that are scaled by the current speed of the head. The table at the bottom contains the complete set of threshold values used.

is generated based on the instantaneously estimated gaze vector and projected until it collides with an object in the virtual environment. This offers an estimation of what stimulus the eyes were foveating at a given time (Watson et al., 2019). Due to the experimental control of VR, location and speed information of each stimulus is known. Combined, this information can be used to identify epochs, where the participant was foveating on either a particular moving or stationary stimulus. To quantify the classification accuracy of our data, we computed the percentage of time points when the eye was foveating a stationary object (the recentering cross or static target) that were erroneously classified as smooth pursuit (including smooth pursuit with compensatory VOR) or head pursuit. This was compared to the percentage of pursuit labeled time points when the eyes were foveating moving objects (i.e., targets in the Dynamic condition). VOR was more difficult to test for as, unlike pursuit, no part of the paradigm necessarily demanded the subject engage in VOR to complete the task. For exploratory purposes, we compared situations when VOR may have been more likely (i.e., just after a saccade to a peripheral static target disk, where the eyes would be left at a more extreme angle and therefore encourage head rotation while maintain gaze on a fixed point) to situations, where VOR may have been less likely (i.e., just after a saccade to a parafoveal static target disk, where perhaps head rotation is less necessary to ensure comfortable gaze position, or during a fixation on the recentering cross).

RESULTS

Simulator Sickness

No subjects experienced symptoms severe enough to withdraw voluntarily from the study. SSQ scores were evaluated using a 2×3 repeated measures ANOVA with time (Pre, Post) and category (Oculomotor, Disorientation, and Nausea) as factors. Greenhouse-Geisser values are reported for the interaction between time and category, which violated sphericity assumptions (Mauchly's $W = 0.741$, $p = 0.037$). Sidak corrections were used for multiple comparisons adjustment. There was a main effect for time $F(1,23) = 14.67$, $p = 0.001$, $\eta^2 = 0.39$ indicating higher average ratings post experiment (12.9, SE 2.5) compared with the start (3.89, SE 1.2) of the experiment. There was a significant main effect for category, $F(2,46) = 11.99$, $p < 0.001$, $\eta^2 = 0.34$, with Oculomotor (12.95, SE 2.4) ratings higher than Disorientation (6.67, SE 1.7) and Nausea (5.57, SE 1.2); both $p < 0.01$. The time by category interaction was also significant $F(2,46) = 5.76$, $p = 0.01$, $\eta^2 = 0.2$ indicating larger post-pre differences in the Oculomotor with respect to the other two categories. An additional paired-samples *t*-test examined the pre/post difference for the Total scores with significantly higher Total scores at the end (15.9, SE 3) compared to the beginning (4.83, SE 1.5) of the experiment $t(23) = 3.8$, $p = 0.001$.

It should be noted that while simulator sickness increased from the start of the experiment to the end, there was no significant difference between saccadic reaction times, $F(1,23) = 0.5$, $p = 0.45$, or button press reaction times, $F(1,23) = 2$,

$p = 0.16$, between the first and last block of the experiment indicating simulator sickness did not have a significant effect on performance.

Eye and Head Responses to Targets at Parafovea and Peripheral Eccentricities

On average, 94% ($SD = 4\%$) of eye data samples were valid in each subject's datafile. This high percentage of valid data can be attributed in part to the eye tracker cameras being embedded inside the headset. This prevents the head and eyes from moving outside the tracking box causing dropped samples, which can occurs in monitor tracking systems when the eyes move outside the confines of the eye tracker (Hessels et al., 2015). This is a feature shared across mobile eye tracking systems more generally (e.g., ETGs and augmented reality devices).

An average of 365 detected saccades ($SE = 28$) were excluded from each participant's data for having an amplitude under 3° and an average of 28 detected saccades ($SE = 2$) were excluded for having a peak velocity over 1,000° deg./s. In total this averaged to 15% of detected saccades being excluded from analysis. Scan paths for the first saccade in each trial are plotted in **Figure 3** as an example of a low (subject 29) and high (subject 22) scan path variability. During the Dynamic condition an average of 1.36 saccades ($SE = 0.06$) were made in parafovea (6° of visual angle from the recentering cross) trials and 1.60 saccades ($SE = 0.09$) in periphery (20° of visual angle from the recentering cross) trials. An average of 1.18 ($SE = 0.04$) saccades were made in parafovea trials and 1.32 ($SE = 0.06$) saccades in periphery trials during the Static condition. The main sequence shown in **Figure 3** exhibits the saccade amplitude by peak velocity relationship for the first saccade in each trial separated by parafoveal and peripheral trials. The mean amplitude of saccades to peripheral cue locations was 17° ($SE = 0.25$) and 18° ($SE = 0.39$) in Static and Dynamic trials, respectively (**Figure 4**). For trials, where the cue appeared in the parafovea the mean saccade amplitude was 6° ($SE = 0.15$) and 6° ($SE = 0.17$) in Static and Dynamic trials, respectively. The skewness and kurtosis were also calculated for Dynamic parafoveal trials ($\gamma = 2.53$, $k = 13.50$), Dynamic peripheral trials ($\gamma = -0.89$, $k = 5.45$), Static parafoveal trials ($\gamma = 1.06$, $k = 5.51$), and Static peripheral trials ($\gamma = -0.15$, $k = 4.01$).

Subjects made a combination of eye and head movements to shift their gaze to the cued target disk (**Figure 5**). Unsurprisingly, larger and faster eye and head movements were made when the cue appeared in the periphery. The time course of head and eye speed shows that, on average, head rotational speed peaks about 200 ms after peak saccadic movement. Together, the head and eye movement measurements in response to the cue onset serve as a quality control check of the eye tracking data collected from HMDs.

Eye Movement Classification

Example trials with classification labels are plotted over eye position in **Figure 6**. VOR classification was rare in this dataset (2% of time points on average) with 1% for parafoveal targets and 3% for peripheral targets. When the ray-cast gaze vector was on the recentering cross, 10% of time points were classified as smooth pursuit in both Static and Dynamic trials. When the gaze vector was on the target, 11% of time points were classified as smooth pursuit in the Static trials compared to 19% in Dynamic trials (where smooth pursuit is likely to occur).

As a follow-up analysis, the average gaze speed was calculated for the longest intersaccadic interval, where the mode ray-cast label of time samples was the target (this was done as some trials contained multiple intersaccadic intervals, where the eyes were foveating on the target). To limit the influence of potential smaller or catch-up saccades on the average gaze speed, time points in which the eye speed exceeded 20 deg./s were excluded from this calculation. This average gaze speed was then plotted against the target's speed relative to the head (**Figure 7**). Plotting the distribution of gaze speeds for each condition, it is apparent that, with a lower gaze threshold, smooth pursuit classification may improve for Dynamic trials in which the target disk was cued in the periphery as the average gaze speed was 9.1 deg./s ($SE = 0.2$ deg./s, $Median = 9.1$ deg./s). However, Dynamic trials in which the target disk was cued in the parafovea elicit a gaze speed ($M = 5.7$ deg./s, $SE = 0.3$ deg./s, $Median = 5.3$ deg./s) that is difficult to isolate from Static parafovea trials ($M = 5.5$ deg./s, $SE = 0.2$ deg./s, $Median = 5.2$ deg./s) and Static peripheral trials ($M = 6.6$ deg./s, $SE = 0.3$ deg./s, $Median = 6.4$ deg./s). Repeating the validation procedure outlined above and separating peripheral and parafoveal targets in the Dynamic conditions shows a smooth pursuit classification of 33 and 10%, respectively (as a reminder 11% of time points were classified as smooth pursuit for static targets). This suggests that the target speed (and associated gaze speed) in Dynamic parafoveal trials was too slow to classify as smooth pursuit using the thresholds in the classification algorithm (**Figure 8**).

Effects of Target Eccentricity and Motion on Task Performance

Saccade and button reaction time to the cue and target, respectively, were analyzed to explore whether target motion or eccentricity affected the speed of subject responses. There was no significant main effect of motion, $F(1,23) = 1.8$, $p = 0.19$, eccentricity, $F(1,23) = 1.7$, $p = 0.20$, or interaction between motion and eccentricity, $F(1,69) = 1.2$, $p = 0.3$, on button press response times. However, first saccades (as defined by the first saccade made after the onset of the target disk cue that measured over 3° in amplitude) were initiated earlier when the target was in the parafovea compared to the periphery, $F(1,23) = 27$, $p < 0.001$ (**Figure 7**). There was no significant effect of motion on saccadic reaction time, $F(1,23) = 3.9$, $p = 0.06$ or interaction between motion and eccentricity, $F(1,69) = 0.7$, $p = 0.4$. We also performed a more conservative analysis, where only trials in which the eyes successfully executed a saccade that went from the recentering cross to the target disk were included. This criterion limited the number of valid trials as subjects often made multiple saccades from the recentering cross to the target disk. As such, 14 subjects had at least 100 trials meeting the criterion and were included in this secondary analysis which

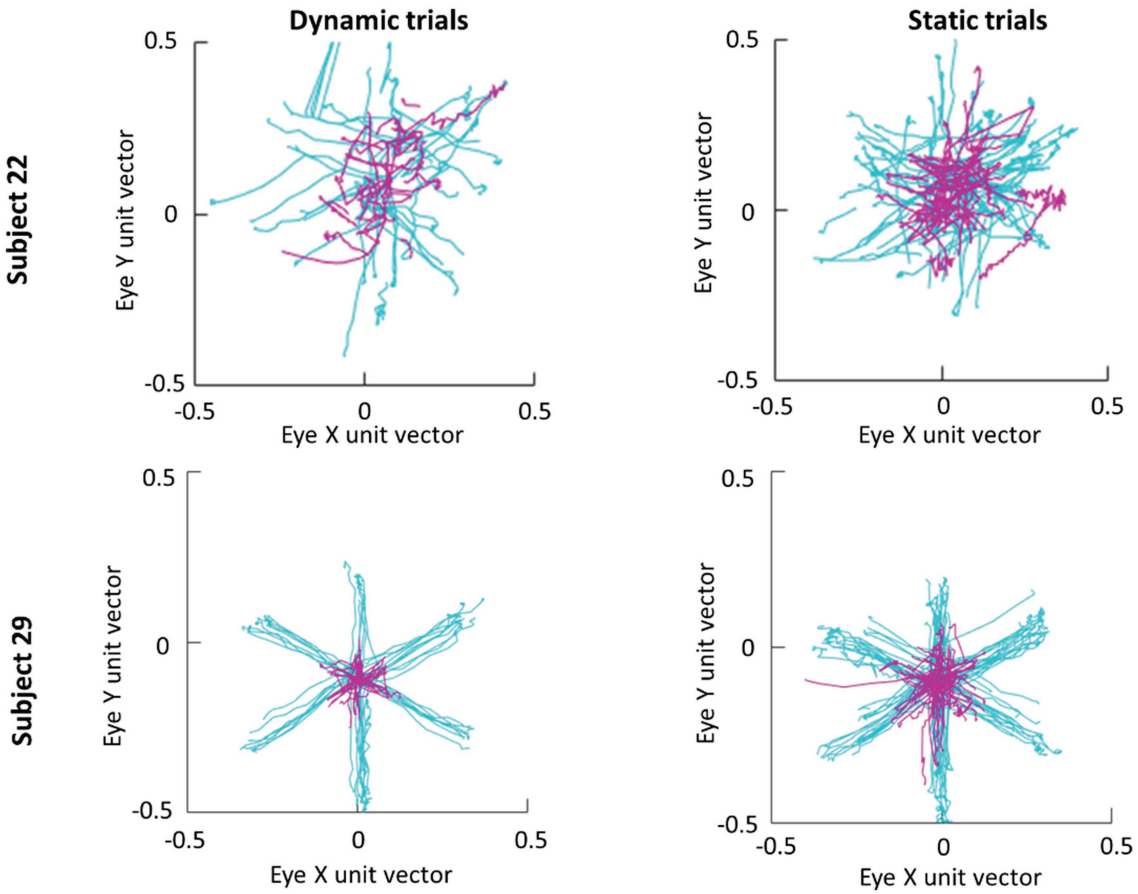


FIGURE 3 | Scan paths of first saccade made in each trial for subjects 22 (top row) and 29 (bottom row). Trials where the cue appeared in the parafovea and periphery are plotted in magenta and cyan, respectively.

also found a significant effect of eccentricity on saccadic reaction times, $F(1,13) = 41$, $p < 0.001$, but not of motion, $F(1,13) = 2.8$, $p = 0.11$, nor any interaction, $F(1,13) = 1.9$, $p = 0.19$. For button reaction times, there was not significant main effects or interactions with the more conservative criterion.

Subjects were instructed to return their gaze back to the recentering cross immediately after button response. This disengagement time, defined as the time between target presentation and the time at which the eyes left the target, was also evaluated. Both eccentricity, $F(1,23) = 87$, $p < 0.001$, and motion, $F(1,23) = 69$, $p < 0.001$ significantly influenced the disengagement time with the eyes leaving dynamic targets and those cued in the periphery earlier (**Figure 8**). There was no significant interaction between motion and eccentricity, $F(1.69) = 0.55$, $p = 0.5$.

DISCUSSION

A fundamental goal of vision researchers is to understand how the human visual system operates in the natural environment. While requirements for experimental control and technological limitations may have necessitated the use of simplified stimuli presented on 2D monitors, the aim has always been to use these results to elucidate mechanisms of real-world vision. While these previous findings have provided an important foundational knowledge, it is essential to ensure that effects seen in the laboratory do in fact translate to the outside world and examine cases in which they do not. For instance, subjects do not exhibit the same detriment in recognition of a scene from a new viewpoint when they themselves have moved to the new viewpoint compared to when the scene is presented in rotated form (as is typical in 2D display experiments; Simons and Wang, 1998). Such findings demonstrate that there are components of natural vision, such as body movement, that are fundamentally integrated with cognition, but are often missed in classic monitor-based experiments. With this in mind, we used a VR HMD with eye tracking technology to study dynamic eye and head movement patterns within a visual discrimination paradigm to induce naturalistic gaze patterns within an immersive yet controlled environment.

The purpose of this work was to demonstrate the capabilities and possible applications of this system as well as to encourage

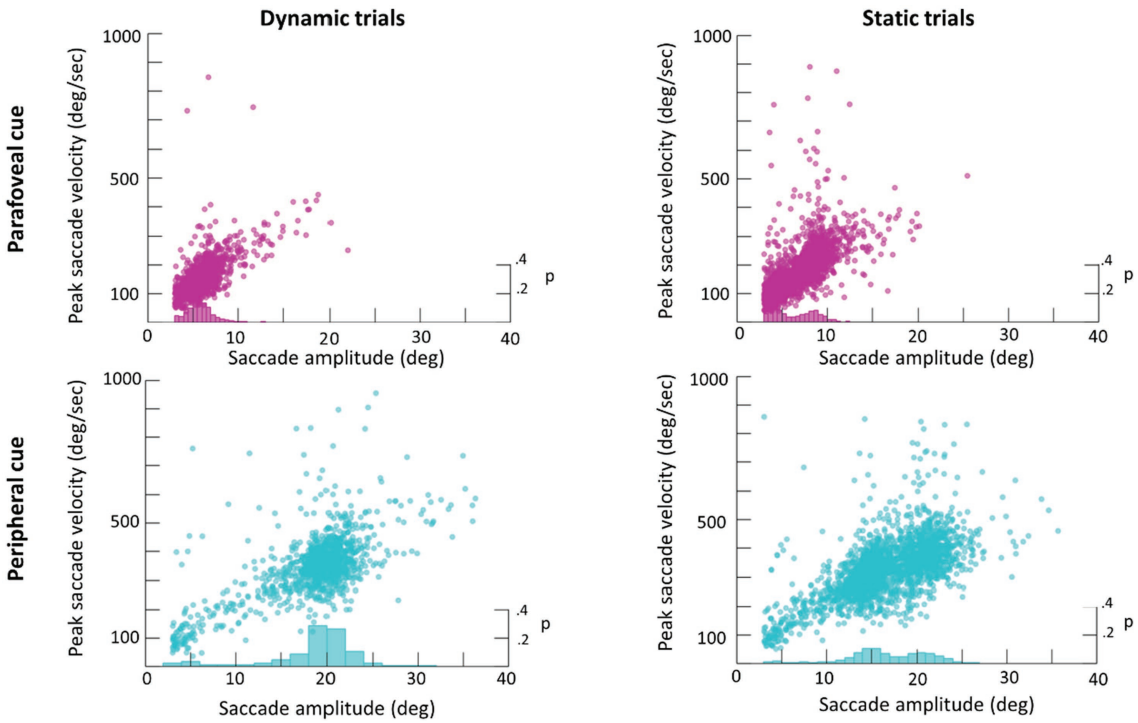


FIGURE 4 | Main sequence scatter plot of saccade amplitude vs. peak amplitude for Dynamic (left) and Static (right) trials. Frequency histograms are plotted on the right y-axis.

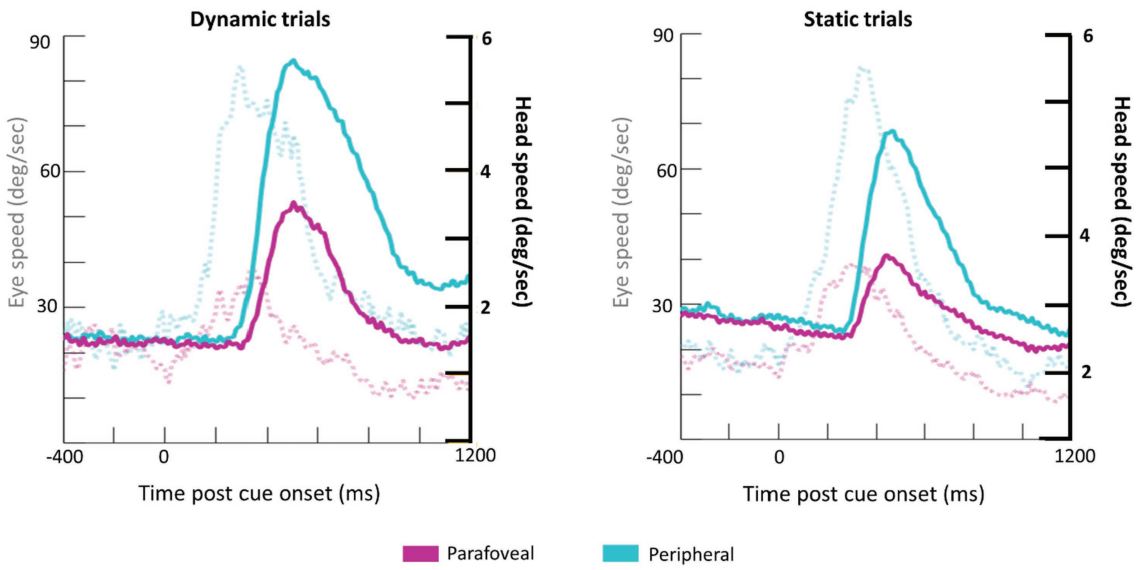


FIGURE 5 | Grand average waveforms for eye speed (transparent, dotted lines; y-axis on the left) and head rotational speed (bold lines; y-axis on the right) for Dynamic (left) and Static (right) trials.

future researchers to incorporate head movements in their exploration of the visual system. To that end, we provided the GUI to lower the bar of entry for vision researchers new to developing paradigms within VR. This GUI allows researchers to quickly and easily set-up a variety paradigms, altering characteristics of the stimuli as well as the relative motion between the participant and the stimulus of background. The intent here is to provide a jumping off point for researchers that may want to move in to the VR with HMD space but hesitate at the upfront programming cost.

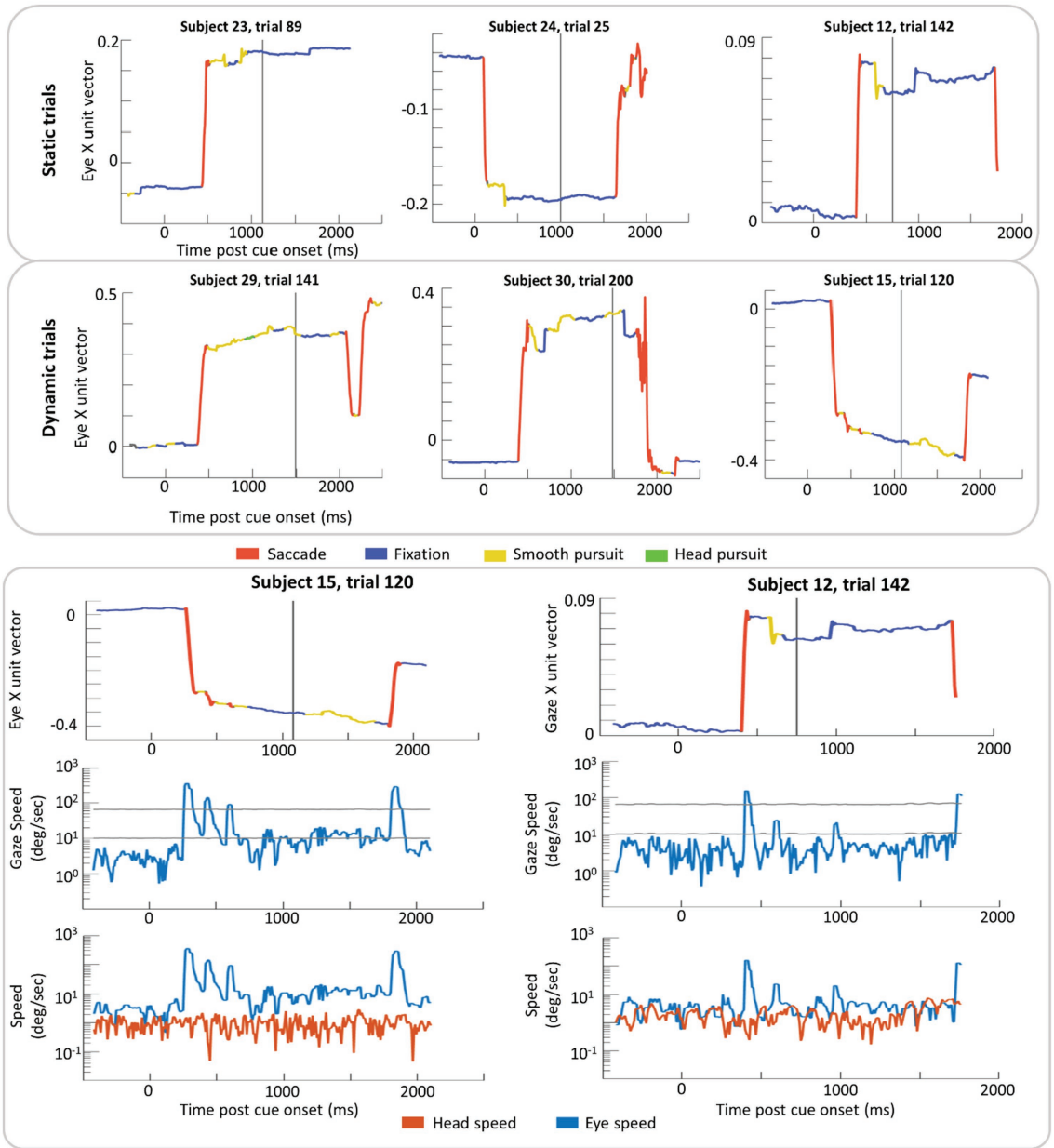


FIGURE 6 | Example trial plots of the eye unit vector along the x-axis for Static and Dynamic trials (top section). In the bottom section, two trials from the top (subject 15, trial 120 and subject 12, trial 142) have been expanded to include the log linear transform of gaze speed plotted with scaled high and low thresholds (grey lines) in addition to the log linear transform of eye and head speed (bottom row). These transformations were done in order to plot instances of high speed without losing detail in slower speed time periods. In each graph, time zero marks the time of the cue (yellow disk) onset and the grey vertical line indicates the time the target was presented. Red lines indicate time windows labeled saccade and blue lines indicate time windows labeled fixation. Yellow and green lines indicate windows in of smooth pursuit and head pursuit, respectively.

The current work used a previously published threshold algorithm designed for use in 360 deg. viewing environments to classify various eye movements such as saccades, fixations, smooth pursuit, and VOR (Agtzidis et al., 2019). To test smooth pursuit classification the current work used the ground truth data about the position of virtual objects to estimate time periods, in which the gaze should be stationary (i.e., when foveating on static objects) compared to when gaze should be engaged in smooth pursuit (i.e., when foveating on moving objects). Separating parafoveal vs. peripheral trials showed that one of the potential contributors to the poor classification accuracy of smooth pursuit was the fact that the relative movement of the disks was too slow to elicit the minimum gaze speed necessary to achieve a "moving" gaze designation. Smooth pursuit classification improved in Dynamic trials, where the target disk was cued in the periphery because targets in the periphery have a higher angular velocity

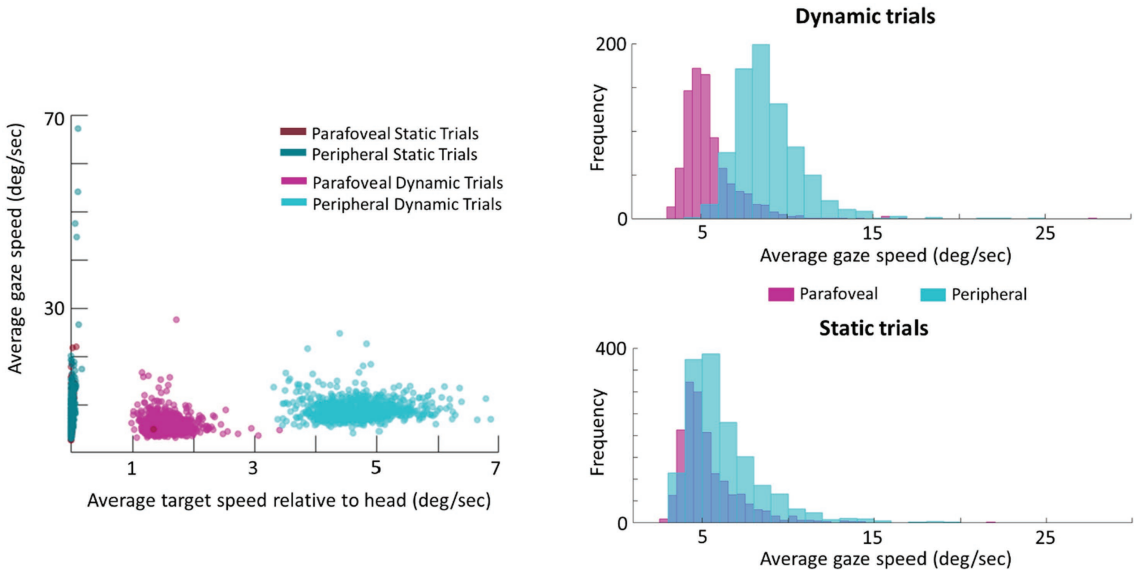


FIGURE 7 | Scatter plot of average gaze speed compared to average target speed (left). Histogram of gaze speed for Dynamic trials (right, top) and Static trials (right, bottom).

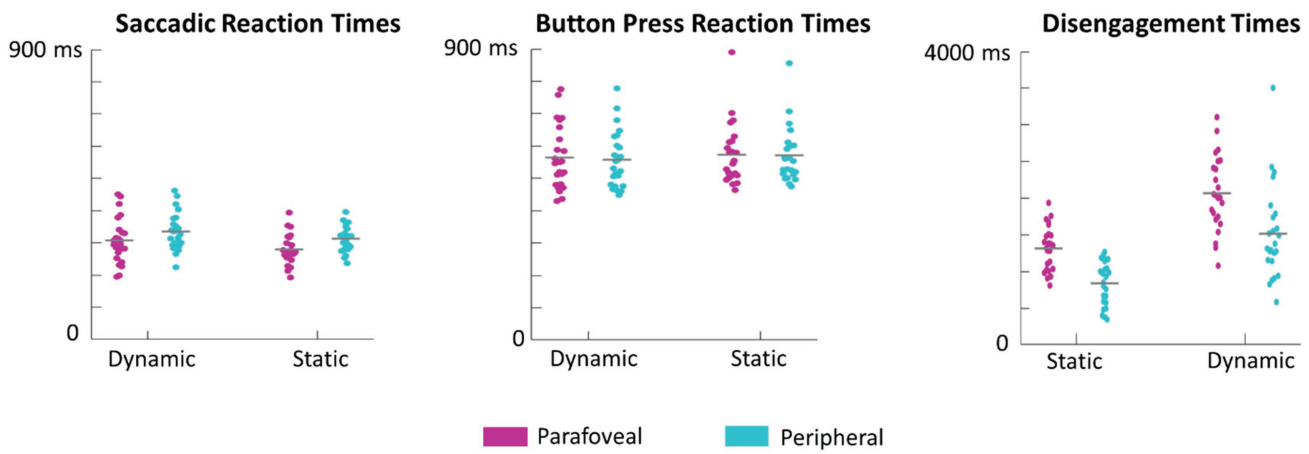


FIGURE 8 | Swarm plots of subject averages in saccadic reaction times, button press reaction times, and disengagement times. It is important to note that saccadic reaction times are calculated as the difference between the time of saccade onset and the time of the cue onset. Button press reaction times are calculated as the difference between when the button was pressed and the time of target onset. Disengagement time is the difference between the time the eyes first left the target and the time of target onset.

relative to the head. Conversely, in Dynamic trials where the target was cued in the parafovea, tracking the target did not elicit gaze speeds that were faster than those of static targets (see an example trial video at https://osf.io/p8g94/). It should be noted that these thresholds were originally set based off an annotated dataset collected by Agtzidis et al. (2016) and therefore could require adjustments based on the speed characteristics of the current stimuli. Smooth pursuit is generally difficult to classify without information about the environment (Agtzidis et al., 2016); however, ideally a classification system does not need to be tailored the specific dataset. Using the directional change of gaze may improve smooth pursuit classification as presumably there would be more coherence in the direction of eye movements when the eyes are tracking a moving stimulus compared to when they are moving around while foveating a static stimulus.

Of course, a limitation of our approach is that it only examines time points in which the eyes are foveating a moving object to test for smooth pursuit, while smooth pursuit can occur without the target object being in the fovea, often necessitating catch-up saccades (de Brouwer et al., 2002). However, using the ray-cast data to isolate when the eyes foveate the target, while not the most conservative approach, should provide a measure of ground truth. It should be noted that on trials where the target was moving faster (Dyanmic, peripheral trials) the accuracy rate

(33%) was comparable to rates found in the original thresholding paper (29–38%; Agtzidis et al., 2019). The results of this technique indicate that strict thresholding is not always sufficient for detecting when the eyes are tracking a moving target at slower speeds. Together, our results demonstrate the complementary benefit of having the ground truth knowledge of stimulus trajectories in VR to similar datasets which use 360° video (David et al., 2018).

Considerations and Limitations of Building Experiments Using HMDs With VR

While there are a number of advantages in using HMD with VR to explore visual processes there are also limitations to consider. For example, given the results obtained using the SSQ, VR researchers should consider ways to mitigate simulation sickness such as smaller fields of view and higher frame rates (e.g., Draper et al., 2001; Keshavarz and Hecht, 2011). Simulator sickness is of special consideration when setting up paradigms meant to elicit specific eye movements. For instance, a potential takeaway from the current findings is that faster moving stimuli should be used in order to reliably classify smooth pursuit. However, this may enhance feelings of simulator sickness and require more frequent breaks or other remedial measures.

Another consideration is that the computer running the game engine will experience fluctuating load levels while rendering the environment. This can be due to visual complexity (e.g., dynamic lighting) or gameplay (e.g., simulation of physical interactor) or other background processes. During periods of low load, the computer can render the environment at sufficiently high frame rates (>100 frames per second). However, as complexity increases, the frame rate can drop substantially depending on the processing capacity of the computer. Importantly, the game engine generates many of the metrics, including position and rotation measures, during each frame refresh. Therefore, the majority of data streams can have a variable sampling rate, not under the direct control of the experimenter, and great care needs to be taken to optimize the simulations performance. It is advisable to set a target frame rate that the simulation will not fall below and to use a computer that is powerful enough to maintain a stable rate. If the environment is not optimized appropriately, rendering can still drop below this target frame rate.

While VR platforms allow for more natural movement, they are restricted in large-scale movement, requiring subjects to teleport themselves through larger virtual environments or incorporate a treadmill. Mobile eye tracking systems avoid this limitation, allowing subjects to navigate the world as they normally would. Another benefit to mobile eye tracking systems is that objects in the environment have real, not simulated, depth, potentially resulting in more natural vergence and accommodation responses.

The VR HMD used here offers eye tracking with a 120 Hz sampling rate. This the low- to mid-range of sampling frequencies necessary to detect and classify eye movements (Holmqvist et al., 2011). The lower sampling rate of the eye trackers in VR may result in less accuracy for measuring small saccades (e.g., microsaccades) and their associated peak velocities. Additionally, lower sampling rates may impair the ability to effectively use certain gaze-contingent interaction, thus preventing adequate online stimulus display changes. Sampling rate of in HMD eye trackers should then be considered not only when designing paradigms but also when comparing results to other eye tracking systems that may have higher sampling rates available.

Another limitation that should be considered is in relating ray-casting to perception. Ray-casting can be a helpful way of labeling an object within foveal vision during a fixation. However, due to noise in the gaze vector estimation and to decreased accuracy compared to desktop eye trackers, it is difficult to accurately classify what object is being foveated if objects in the environment are too close to one another. Watson et al. (2019) offers an alternative to this approach with a "shotgun" ray-cast that returns a list of objects contained in the area surrounding gaze position. However importantly, neither a pin-point or shotgun ray-cast gives explicit insight into what objects in the visual field are actually attended to or encoded into memory and this should be kept in mind when drawing conclusions of perception from ray-cast data.

Lastly, the HMD used here utilizes "Outside In" tracking, requiring external lighthouses containing infrared scanners to be mounted in opposing corners of the tracking area. These lighthouses contain spinning mirrors, and so are susceptible to vibrations if not firmly mounted. Also, reflective surfaces could potentially disrupt the headset's ability to track the lighthouses. It is important to reduce reflective surfaces and firmly mount lighthouses when using a headset with Outside In tracking. Additionally, newer versions of this technology, such as the Tobii Vive Pro, allow for the installation of two additional lighthouses (which is the maximum number available with the system used here) which may help to address tracking issues (Niehorster et al., 2017).

CONCLUSION

Virtual reality used in conjunction with HMD offers a potential solution to vision researchers by offering a balance between allowing more naturalistic behavior in participants without sacrificing strict experimental control. Here, we offer a demonstration of some of the capabilities of this system as well as the GUI for future researchers to be able to quickly launch a variety of visual search paradigms to suit research needs. There are a number of technological hurdles to consider when experimenting within VR which we have outlined above. However, overall this technology offers a promising space for understanding how vision is performed in the natural environment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ARL Human Research Protection Program. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CB developed all Unity code for the paradigm and its GUI as well as collected data. AR came up with the idea and designed the experimental task and collected data. CC-F performed the analysis and wrote the manuscript. JT provided support for the creation of the paradigm as well as subsequent data analysis. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.650693/full#supplementary-material

REFERENCES

Agtzidis, I., Meyhöfer, I., Dorr, M., and Lencer, R. (2020). Following forrest gump: smooth pursuit related brain activation during free movie viewing. *NeuroImage* 216:116491. doi: 10.1016/j.neuroimage.2019.116491

Agtzidis, I., Startsev, M., and Dorr, M. (2016). Smooth pursuit detection based on multiple observers. *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking and Applications*; March 14–17, 2016. 303–306.

Agtzidis, I., Startsev, M., and Dorr, M. (2019). 360-degree video gaze behavior: A ground-truth data set and a classification algorithm for eye movements. *Proceedings of the 27th ACM International Conference on Multimedia;* October 21–25, 2019. 1007–1015.

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., and Donk, M. (2015). It depends on when you look at it: salience influences eye movements in natural scene viewing and search early in time. *J. Vis.* 15:9. doi: 10.1167/15.5.9

Baloh, R. W., Konrad, H. R., Sills, A. W., and Honrubia, V. (1975). The saccade velocity test. *Neurology* 25:1071. doi: 10.1212/WNL.25.11.1071

Boettcher, S. E., Draschkow, D., Dienhart, E., and Vö, M. L. H. (2018). Anchoring visual search in scenes: assessing the role of anchor objects on eye movements during visual search. *J. Vis.* 18:11. doi: 10.1167/18.13.11

Clay, V., König, P., and König, S. U. (2019). Eye tracking in virtual reality. *J. Eye Mov. Res.* 12, 1–18. doi: 10.16910/jemr.12.1.3

David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., and Callet, P. L. (2018). A dataset of head and eye movements for 360 videos. *Proceedings of the 9th ACM Multimedia Systems Conference;* June 12–18, 2018. 432–437.

de Brouwer, S., Missal, M., Barnes, G. R., and Lefèvre, P. (2002). Quantitative analysis of catch-up saccades during sustained pursuit. *J. Neurophysiol.* 87, 1772–1780. doi: 10.1152/jn.00621.2001

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *J. Exp. Psychol. Gen.* 140, 552–572. doi: 10.1037/a0023885

Dorr, M., Martinetz, T., Gegenfurtner, K. R., and Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *J. Vis.* 10:28. doi: 10.1167/10.10.28

Draper, M. H., Viirre, E. S., Furness, T. A., and Gawron, V. J. (2001). Effects of image scale and system time delay on simulator sickness within head-coupled virtual environments. *Hum. Factors* 43, 129–146.

Engbert, R., and Kiegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vis. Res.* 43, 1035–1045. doi: 10.1016/S0042-6989(03)00084-1

Engbert, R., and Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7192–7197. doi: 10.1073/pnas.0509557103

Harada, Y., and Ohyama, J. (2019). Spatiotemporal characteristics of 360-degree basic attention. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-52313-3

Haskins, A. J., Mentch, J., Botch, T. L., and Robertson, C. E. (2020). Active vision in immersive, 360 real-world environments. *Sci. Rep.* 10, 1–11. doi: 10.1038/s41598-020-71125-4

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., and Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye Movements: A Window on Mind and Brain*. (eds.) GompelR. P. G. van, M. H. Fischer, W. S. Murray and R. L. Hill. (Oxford: Elsevier), 537–562.

Hessels, R. S., Cornelissen, T. H., Kemner, C., and Hooge, I. T. (2015). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behav. Res. Methods* 47, 848–859. doi: 10.3758/s13428-014-0507-6

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*. Oxford: Oxford University Press.

Jangraw, D. C., Johri, A., Gribetz, M., and Sajda, P. (2014). NEDE: an open-source scripting suite for developing experiments in 3D virtual environments. *J. Neurosci. Methods* 235, 245–251. doi: 10.1016/j.jneumeth.2014.06.033

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220. doi: 10.1207/s15327108ijap0303_3

Keshavarz, B., and Hecht, H. (2011). Validating an efficient method to quantify motion sickness. *Hum. Factors* 53, 415–426.

Kim, H. C., Jin, S., Jo, S., and Lee, J. H. (2020). A naturalistic viewing paradigm using 360° panoramic video clips and real-time field-of-view changes with eye-gaze tracking. *NeuroImage* 216:116617. doi: 10.1016/j.neuroimage.2020.116617

Kothe, C. (2014). Lab streaming layer (LSL).

Niehorster, D. C., Li, L., and Lappe, M. (2017). The accuracy and precision of position and orientation tracking in the HTC vive virtual reality system for scientific research. *Iperception* 8:2041669517708205. doi: 10.1177/2041669517708205

O'Connell, T. P., and Chun, M. M. (2018). Predicting eye movements patterns from fMRI responses to natural scenes. *Nat. Commun.* 9, 1–15. doi: 10.1038/s41467-018-07471-9

Rai, Y., Gutiérrez, J., and Le Callet, P. (2017). A dataset of head and eye movements for 360 degree images. *Proceedings of the 8th ACM International Conference on Multimedia;* June 20, 2017. 205–210.

Ries, A. J., Slayback, D., and Touryan, J. (2018). The fixation-related lambda response: effects of saccade magnitude, spatial frequency, and ocular artifact removal. *Int. J. Psychophysiol.* 134, 1–8. doi: 10.1016/j.ijpsycho.2018.09.004

Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A., Rizzo, A., and Loetscher, T. (2020). Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. *Front. Hum. Neurosci.* 14:96. doi: 10.3389/fnhum.2020.00096

Simons, D. J., and Wang, R. F. (1998). Perceiving real-world viewpoint changes. *Psychol. Sci.* 9, 315–320. doi: 10.1111/1467-9280.00062

Terao, Y., Fukuda, H., and Hikosaka, O. (2017). What do eye movements tell us about patients with neurological disorders?-An introduction to saccade recording in the clinical settings. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 10, 772–801. doi: 10.2183/pjab.93.049

Ward, L. M., and Kapoula, Z. (2020). Differential diagnosis of vergence and saccade disorders in dyslexia. *Sci. Rep.* 10, 1–15. doi: 10.1038/s41598-020-79089-1

Watson, M. R., Voloh, B., Thomas, C., Hasan, A., and Womelsdorf, T. (2019). USE: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificial intelligent agents. *J. Neurosci. Methods* 326:108374. doi: 10.1016/j.jneumeth.2019.108374

Williams, C. C., and Castelhano, M. S. (2019). The changing landscape: high-level influences on eye movement guidance in scenes. *Vision* 3:33. doi: 10.3390/vision3030033

Wolfe, J., Võ, M. L. H., Evans, K. K., and Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends Cogn. Sci.* 15, 77–84. doi: 10.1016/j.tics.2010.12.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 February 2022
doi: 10.3389/fpsyg.2022.788231

Spontaneous Eye Blink Rate During the Working Memory Delay Period Predicts Task Accuracy

*Jefferson Ortega[1], Chelsea Reichert Plaska[1,2], Bernard A. Gomes[3] and Timothy M. Ellmore[1,2]**

[1]*Department of Psychology, The City College of the City University of New York, New York, NY, United States,* [2]*Behavioral and Cognitive Neuroscience Program, The Graduate Center of the City University of New York, New York, NY, United States,* [3]*Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA, United States*

Spontaneous eye blink rate (sEBR) has been linked to attention and memory, specifically working memory (WM). sEBR is also related to striatal dopamine (DA) activity with schizophrenia and Parkinson's disease showing increases and decreases, respectively, in sEBR. A weakness of past studies of sEBR and WM is that correlations have been reported using blink rates taken at baseline either before or after performance of the tasks used to assess WM. The goal of the present study was to understand how fluctuations in sEBR during different phases of a visual WM task predict task accuracy. In two experiments, with recordings of sEBR collected inside and outside of a magnetic resonance imaging bore, we observed sEBR to be positively correlated with WM task accuracy during the WM delay period. We also found task-related modulation of sEBR, including higher sEBR during the delay period compared to rest, and lower sEBR during task phases (e.g., stimulus encoding) that place demands on visual attention. These results provide further evidence that sEBR could be an important predictor of WM task performance with the changes during the delay period suggesting a role in WM maintenance. The relationship of sEBR to DA activity and WM maintenance is discussed.

Keywords: spontaneous eye blink rate, working memory, delay period, dopamine, attention

OPEN ACCESS

Edited by:
Steven Matthew Thurman, United States Army Research Laboratory, United States

Reviewed by:
Stephen B. R. E. Brown, Red Deer Polytechnic, Canada
Matthew Robison, University of Texas at Arlington, United States

****Correspondence:***
Timothy M. Ellmore
tellmore@ccny.cuny.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *01 October 2021*
Accepted: *11 January 2022*
Published: *15 February 2022*

Citation:
Ortega J, Plaska CR, Gomes BA and Ellmore TM (2022) Spontaneous Eye Blink Rate During the Working Memory Delay Period Predicts Task Accuracy. Front. Psychol. 13:788231. doi: 10.3389/fpsyg.2022.788231

INTRODUCTION

The healthy human blinks around 15–20 times per minute (Tsubota et al., 1996), however the precorneal tear film, which lubricates the eye, only begins drying up approximately 25 s after a blink ends (Norn, 1969). This suggests that we blink more often than needed to maintain a lubricated precorneal tear film. Previous research has found task-related modulation of spontaneous eye blink rate (sEBR), which indicates that blinking could be reflective of cognitive factors (Siegle et al., 2008; Oh et al., 2012). For example, reading is accompanied by low levels of sEBR, while high levels of sEBR have been reported during conversation (Bentivoglio et al., 1997). More recent studies have found that sEBR correlates with attentional load and fatigue (Maffei and Angrilli, 2018), attentional control (Colzato et al., 2009; Unsworth et al., 2019a), can track working memory updating and gating (Rac-Lubashevsky et al., 2017), and can predict differences in exploration during reinforcement learning (Van Slooten et al., 2019). In addition, a growing body of literature continues to provide evidence supporting

sEBR as an effective measure of striatal dopamine (DA) activity (Jongkees and Colzato, 2016). However, whether sEBR does indeed reflect DA activity is still debated today (Dang et al., 2017; Sescousse et al., 2018).

The connection between sEBR and DA first came from observations of neurological and psychiatric disorders that found decreased sEBR in patients with Parkinson's (Hall, 1945; Reddy et al., 2013), a neurodegenerative disorder that affects the dopaminergic system in the brain, causing symptoms like rigidity (Dauer and Przedborski, 2003). Schizophrenia has also been suggested to provide evidence for a connection between sEBR and DA due to excessive DA activity in the striatum (Howes et al., 2015) and increased sEBR in schizophrenia patients (Adamson, 1995; Swarztrauber and Fujikawa, 1998). Additionally, sEBR and DA has previously been investigated in pharmacological studies, which have observed an increase in sEBR after administration of DA agonists, while DA antagonists decreased sEBR in primates (Elsworth et al., 1991; Jutkiewicz and Bergman, 2004). In one study, researchers found sEBR was correlated with dopamine levels specifically in the caudate nucleus in monkeys, suggesting that DA could regulate blink rate (Taylor et al., 1999). This is further supported by another study that found sEBR to be closely related to *in vivo* and positron emission tomography measures of striatal D2 receptor density in the ventral striatum and caudate nucleus of adult male vervet monkeys (Groman et al., 2014). These findings provide valuable evidence for sEBR being a viable measure of striatal DA activity and have led to many researchers to adopt sEBR as a measure of DA activity. Moreover, sEBR is an easy-to-record physiological measure that is non-invasive and affordable.

One cognitive process of interest, that is also closely related to DA activity, is working memory (WM) which is the process of actively holding information online and manipulating it to meet task demands (Baddeley, 1992). Prior research has found substantial evidence that demonstrates the importance of dopaminergic neurotransmission and the role of the prefrontal cortex during WM function (Fuster and Alexander, 1971; Funahashi et al., 1989; Courtney et al., 1998; Wager and Smith, 2003; Cools and Robbins, 2004), especially during WM maintenance (Fuster and Alexander, 1971; Funahashi et al., 1989; Constantinidis et al., 2018). Specifically, human studies investigating DA in WM tasks have found both caudate dopamine activity during WM maintenance and DA synthesis capacity to be positively correlated with WM capacity, a measure of the amount of information that can be held in WM (Cools et al., 2008; Landau et al., 2009). Though it is widely accepted that the PFC plays an important role in WM function (Roberts et al., 1998), many researchers still debate the PFC's role in WM (Seamans and Yang, 2004). One model that attempts to elucidate the PFC's role in WM function is the prefrontal cortex basal ganglia WM model (PBWM; Frank et al., 2001; Hazy et al., 2006). PBWN is a computational neural network model that suggests that WM requires robust maintenance and rapid selective updating. This model states that the frontal cortex facilitates robust, active maintenance through recurrent excitation in frontal neurons, while the basal ganglia orchestrates a gating mechanism that controls the flow of information into WM (Frank et al., 2001). Previous research has pointed toward DA being important for this sustained firing activity in the PFC during WM maintenance (Sawaguchi, 2001; Durstewitz and Seamans, 2008; De Frias et al., 2010). The relationship between DA and WM performance is believed to follow an inverted U-shape, in which too little or too much dopamine impairs performance, as seen in psychopharmacological studies (Stewart and Plenz, 2006). In one study, the effects of administered dopaminergic drugs on PFC function depended on baseline levels of performance, whereas administration of bromocriptine, a dopamine agonist, impaired performance for individuals with higher working memory abilities while improving performance for individuals with lower working memory abilities (Kimberg et al., 1997).

Although sEBR has been used in prior research to investigate cognitive functions like WM, many of these studies relied on baseline levels of sEBR to investigate these relationships (Tharp and Pickering, 2011; Zhang et al., 2015; Unsworth et al., 2019b). Few studies have investigated the relationship between phasic sEBR during a WM task. Phasic sEBR refers to the measuring of sEBR in response to stimulus conditions while tonic sEBR refers to baseline levels of blinking (Bacher and Allen, 2009). To the best of our knowledge, only one other study has examined sEBR as a function of the different task phases (e.g., stimulus encoding, maintenance during the delay period, and stimulus probe periods) of a WM task (Bacher et al., 2017). Bacher et al. (2017) found modulation of sEBR across these different phases are developed in infants as young as 10 months, indicating that sEBR can reflect dopamine function in early human development. They also observed higher sEBR during the Hide (delay) phase of the task in relation to the Reveal phase, which is when the experimenter revealed the toy's location to the child. This modulation of sEBR was suggested to reflect the engagement of cognitive resources that have become available during the Hide phase and transiently elevated DA activity that is needed to update and maintain mental representations (Bacher et al., 2017).

The goal of the current study was to investigate how fluctuations in sEBR during different phases of a Sternberg visual WM task (**Figure 1**) relate to performance, and how sEBR fluctuations change across task demands. First, we hypothesized that sEBR during the WM Delay period, when stimuli are being maintained, would be positively correlated with task performance and that there would be a non-linear relationship such that low and high sEBR would correlate with worse performance. Second, we hypothesized differences in sEBR across phases of the WM task with differences between phasic sEBR during the WM delay and tonic sEBR during non-task rest periods.

MATERIALS AND METHODS

Participants

The experiments were conducted under a protocol approved by the Institutional Review Board of the City University of New York Human Research Protection Program (CUNY HRPP IRB).

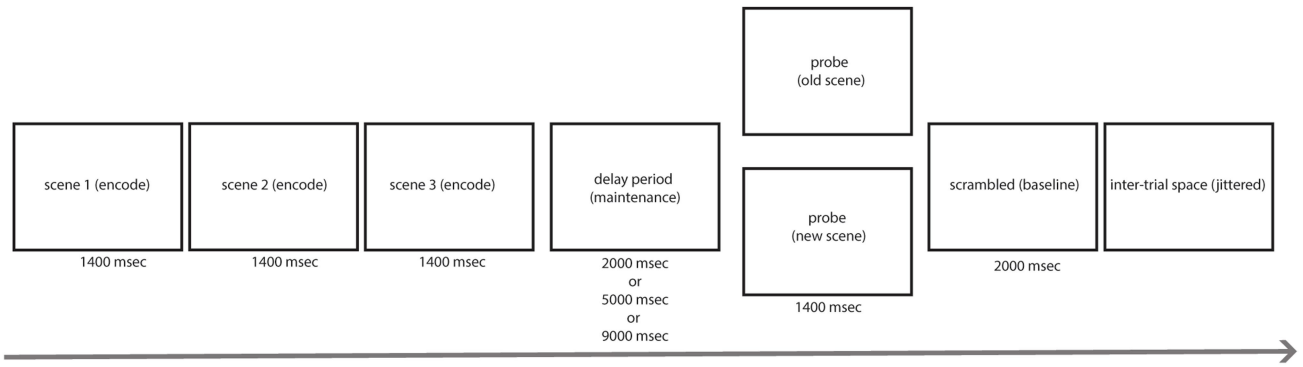


FIGURE 1 | Task design. Each trial began with an encoding period where three novel complex scenes were presented for 1,400 ms each. The encoding period was followed by a delay period where a fixation cross was presented on a gray background for a varied amount of time (2 s, 5 s, or 9 s). After the delay period, the probe was presented for 1,400 ms and participants had to identify whether the image was a new image or one of the previously presented images with a button press. After the probe, a scrambled image was presented for 2,000 ms which indicated the end of the trial followed by jittered blank space before the start of the next trial.

All methods were carried out in accordance with the relevant guidelines and regulations of the CUNY HRPP IRB committee. All participants were recruited either by flyers posted throughout the City College of New York campus or by web postings on the City College of New York SONA online experimental scheduling system. All participants had normal or corrected-to-normal vision with no reported neurological or psychiatric disorders. Participants were either compensated $15 per hour or received one psychology course credit per hour of participation in the study. Written informed consent was obtained from all participants in the study.

Participants selected for Experiment 1 and Experiment 2 were part of a larger study. Nineteen healthy participants (8 males; $M = 23.79$; $SD = 7.72$) were recruited for Experiment 1. In Experiment 1, sEBR was measured inside a 3 tesla Siemens Prisma MRI scanner. In Experiment 2, sEBR was recorded in a sound attenuated EEG booth during acquisition of EEG data while participants sat upright. Fifty-three healthy participants (29 males; $M = 23.58$; $SD = 5.79$) were recruited for Experiment 2. Three participants were removed from Experiment 1 including one participant who was removed for noisy data and two who were removed for task performance below or close to chance. A total of 19 participants were removed from Experiment 2 for multiple reasons including 11 participants who were removed due to bad EOG channel quality, four participants who were removed because of a stimulus marker malfunction, three participants who were removed due to outlier detection, and two who were removed for failing to adhere to the protocol. The final sample for the analysis in Experiment 1 was 16 subjects, and for Experiment 2 was 34 subjects.

Task and Procedure

Prior to the start of the task, participants completed a 5-min Rest period which consisted of staring at a black fixation cross that was shown on a gray background. Participants completed another 5-min Rest period after completing three runs of the task. This fixation cross was also used during the delay period of the task. Participants completed three runs, each run containing 54 trials, of a modified version of the Sternberg WM task (Sternberg, 1966). Naturalistic scenes were used as stimuli and were sampled from the SUN database (Xiao et al., 2010). The task consisted of a stimulus encoding period, delay period, probe period, and post-probe scrambled stimulus period (which served as a visual baseline and to signal end of trial). During the encoding period, participants were shown three subsequent novel scenes for 1400 ms each. During the delay period, a black fixation cross was shown on a gray background for varied lengths (either 2, 5, or 9 s long). The delay period duration was randomized from trial to trial to engage subjects' attention consistently across trials because they could not predict when the delay period would end. Each three runs of the task had 18 trials of each delay duration with order of presentation randomized. The probe was presented for 1400 ms after the delay period and consisted of a new image (one that has not been presented yet) or an old image (one that was shown during encoding). The chance of receiving a new probe was 50%. Participants indicated whether the image presented was either a new or an old image with a button press. After the probe, a Fourier phase-scrambled scene was shown for 2000 ms, indicating the end of the current trial followed by a jittered period of blank screen.

sEBR Recording

Participants were not given instructions about when to blink during the experiment. Previous studies have found blink rate to be stable between 10 am and 5 pm (Barbato et al., 2000; Doughty and Naase, 2006). For both Experiment 1 and Experiment 2, sEBR was recorded between 10:00 am and 3:00 pm. During Experiment 1, eye blinks were recorded inside a three tesla Siemens Prisma MRI scanner using an MRI compatible EyeLink 1,000 Eye Tracker (SR Research) and was recorded at 500 Hz. In Experiment 2, eye blinks were recorded using an electrooculogram (EOG) that was recorded during 64-channel

scalp electroencephalogram using a Brain Products cap and active electrode recording system. EOGs were placed above the left eye and below the right eye to track blinking. Blink detection was performed using MNE Python *via* the function "*find_eog_events*" (Gramfort et al., 2013). Blink epochs were evaluated for each run of the task for all participants. Runs with blink epochs which did not resemble the standard blink shape were removed from the analysis. Only participants with 2 or more runs of good eye-tracking data were used in the analysis. The first 2 s of all delay periods were used in the analysis. In Experiment 1 and 2, sEBR was computed by dividing the total number of blinks by the total period duration for any given phase resulting in units of blinks per minute:

$$sEBR = \frac{total\ blinks}{total\ period\ duration}$$

Statistical Analysis

Statistical analyses were computed using JASP Version 0.16. sEBR and task accuracy data were checked for outliers prior to analysis and were removed. Because the relationship between sEBR and WM performance is believed to follow an "inverted U-shape," we did not consider Pearson's r the optimal measure for this analysis because it is limited to evaluating only a linear relationship between two variables. Instead, we computed Spearman's rho, which can describe monotonic functions, whereas the value of one variable changes the other variable changes but not necessarily at a constant rate. We also used polynomial regression analysis, which is more appropriate for quantifying non-linear associations. Specifically, we investigated the non-linear relationship between task accuracy and sEBR during the Delay period of the task for both Experiment 1 and Experiment 2 using a second order polynomial regression model. Unidimensional reliability analyses were computed using task accuracy and sEBR as input variables and Cronbach's α as the frequentist scale reliability statistic. Post-hoc statistical power calculations were computed for each experiment and the combined samples of both experiments with G*Power Version 3.1.9.6 using the correlation between sEBR during the Delay period and task accuracy. Parameters included the Exact test family, Correlation: Bivariate normal model with an α error probability of 0.05, the sample size, and the correlation coefficient as the effect size.

RESULTS

Experiment 1

In Experiment 1, we examined the relationship between sEBR and WM task performance while the duration of the WM delay period interval was varied. The first 2 s of all Delay periods were used in the analysis. First, because of the previously reported non-linear relationship between DA and WM task performance (Cools and D'Esposito, 2011), Spearman's rho correlation coefficient (r_s) was used to analyze the relationship between sEBR and task accuracy. After performing Bonferroni multiple comparisons correction on values of p, we found no significant relationship between sEBR during the phases of the task and task accuracy (**Figure 2A**). However, there was a strong positive correlation between sEBR during the Delay period and task accuracy ($r_s = 0.526$, $p = 0.036$; **Figure 2A**).

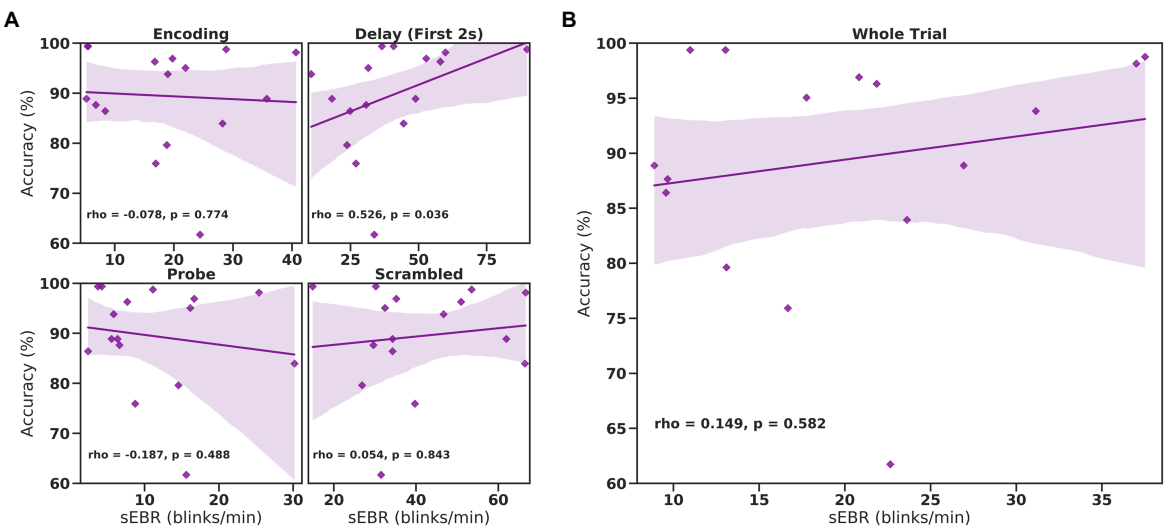


FIGURE 2 | Correlation between sEBR during different phases of the task and task accuracy in Experiment 1. Correlation plots show sEBR (blinks/min) on the x-axis and task accuracy on the y-axis. **(A)** These four plots are encoding (top left), the first 2 s of the delay (top right), probe (bottom left), and scrambled (bottom right) periods. The delay period shows a positive correlation ($p = 0.036$ but not significant after multiple comparisons correction) between task accuracy and sEBR during the first 2 s of the delay period. **(B)** This plot represents the relationship between sEBR during the whole trial and task accuracy. Fitted line represents linear regression model fit. Shaded region depicts 95% confidence interval.

We then examined the correlation between sEBR during the whole trial period and task accuracy to make sure that this relationship was not driving the relationship between sEBR during the Delay and task accuracy. There was no significant relationship between sEBR during the whole trial and task accuracy ($r_s = 0.149$, $p = 0.582$; **Figure 2B**). Descriptive statistics and reliability measures for Experiment 1 are presented in **Table 1**. Second, we computed a repeated measures ANOVA test to compare participants' sEBR across the task phases. A Mauchly's test of sphericity was first computed to check the assumption of sphericity in the data and was found to be significant ($p = 0.012$). Greenhouse-Geisser and Huynh-Feldt ε values were smaller than 0.75 so a Greenhouse-Geisser correction was performed. There were significant differences in sEBR between group means [F (1.948, 29.213) = 33.196, $p < 0.001$]. A *post-hoc* test using the Holm correction revealed that sEBR was significantly lower during Encoding (18.9 ± 11.0 sEBR, $p < 0.001$) and Probe (11.3 ± 8.0 sEBR, $p < 0.001$) periods compared to the Delay period (39.4 ± 19.5 sEBR; **Figure 3**). There was no significant difference in sEBR between the Delay and Scrambled period ($p = 0.682$). sEBR was also significantly lower during Encoding (18.9 ± 11.0 sEBR, $p < 0.001$) and Probe (11.3 ± 8.0 sEBR, $p < 0.001$) periods compared to the Scrambled period (40.9 ± 15.2 sEBR; **Figure 3**). Finally, we investigated the difference between phasic sEBR during the Delay period and tonic sEBR during the Rest period. We performed a paired samples T test to compare sEBR during the Delay and during Rest. We observed sEBR to be significantly higher during the Delay period (39.4 ± 19.5 sEBR) compared to the Rest period (28.6 ± 14.7 sEBR), $t(15) = 2.885$, $p = 0.0011$ (**Figure 4A**). We then investigated the correlation between tonic sEBR during the Rest period and task accuracy. There was no significant correlation between sEBR during the Rest period and task accuracy ($r_s = 0.259$, $p = 0.333$; **Figure 4B**).

Experiment 2

Experiment 2 included a larger sample of subjects with a task design identical to Experiment 1. First, we examined the relationship between sEBR during each WM task phase and task accuracy. After performing Bonferroni correction on values of p, we found that sEBR during the WM delay period was correlated positively with task performance ($r_s = 0.508$, $p = 0.002$), with no significant relationships observed between sEBR in other task periods and task performance (**Figure 5A**). We then examined the relationship between sEBR during the whole trial and task accuracy to make sure that the significant relationship between Delay sEBR and task accuracy was not driven by sEBR during the whole trial. We found no significant relationship between whole trial sEBR and task accuracy ($r_s = 0.192$, $p = 0.278$; **Figure 5B**). Descriptive statistics and reliability measures for Experiment 2 are presented in **Table 2**. We then repeated the same analysis of comparing sEBR across the task phases by computing a repeated measures ANOVA test. A Mauchly's test of sphericity was first computed to check the assumption of sphericity in the data and was found to be significant ($p < 0.001$). Greenhouse-Geisser and the Huynh-Feldt ε values were smaller than 0.75 so a Greenhouse-Geisser correction was performed. There were significant differences in sEBR between group means [F (1.578,52.058) = 66.958, $p < 0.001$]. A *post-hoc* test using the Holm correction revealed that sEBR was significantly lower during Encoding (11.6 ± 8.0 sEBR, $p < 0.001$), Probe (7.3 ± 4.1 sEBR, $p < 0.001$), and Scrambled (19.5 ± 10.6 sEBR, $p < 0.001$) periods compared to the Delay period (35.6 ± 18.3 sEBR; **Figure 6**). sEBR was also significantly lower during the Encoding (11.6 ± 8.0 sEBR, $p < 0.001$) and Probe (7.3 ± 4.1 sEBR, $p < 0.001$), periods compared to the Scrambled period (19.5 ± 10.6 sEBR; **Figure 6**). Additionally, sEBR was significantly lower during the Probe (7.3 ± 4.1 sEBR, $p = 0.047$), period compared to the Encoding period (11.6 ± 8.0 sEBR; **Figure 6**). We then investigated the difference between sEBR during the Delay period and sEBR during the Rest period. We performed a paired samples T test to compare sEBR during the Delay and during Rest. We observed sEBR to be significantly higher during the Delay period (35.6 ± 18.3 sEBR) compared to the Rest period (17.7 ± 11.1 sEBR), $t(33) = 6.005$, $p < 0.001$ (**Figure 7A**). We then investigated the correlation between tonic sEBR during the Rest period and task accuracy. There was no significant correlation between sEBR during the Rest period and task performance ($r_s = -0.053$, $p = 0.768$; **Figure 7B**).

Polynomial Regression Model

To investigate whether sEBR during the Delay varies non-linearly with task performance, we computed a quadratic polynomial regression model between sEBR during the Delay period of Experiment 1 and Experiment 2 and task accuracy. There was no significant polynomial regression relationship found between

TABLE 1 | Descriptive statistics and split-half reliability for Experiment 1.

Variable	*n*	*M*	*SD*	Skewness	Kurtosis	Split-half coefficient
Task Accuracy (%)	16	89.43	10.34	−1.41	2.15	0.944
Whole Trial sEBR	16	20.07	9.40	0.62	−0.60	0.906
Encoding sEBR	16	18.89	10.95	0.39	−0.54	0.965
Delay sEBR	16	39.44	19.45	1.06	1.73	0.974
Probe sEBR	16	11.29	7.97	1.17	0.83	0.903
Scrambled sEBR	16	40.94	15.18	0.45	−0.62	0.961
Rest sEBR	16	28.59	14.68	1.05	0.73	

M and SD represent mean and standard deviation, respectively. Split-half reliability based on 1,000 bootstrap replicates.

task accuracy and the first 2 s of the Delay in Experiment 1 ($\beta = 0.324$, $p = 0.741$) nor between task accuracy and the first 2 s of the Delay in Experiment 2 ($\beta = -0.568$, $p = 0.322$; **Figure 8**).

Reliability and Statistical Power

Reliability of each measure was computed using a split-half analysis procedure. Each measure was divided into two subsets, at random, by trial and recomputed. Correlation coefficients were then calculated on both subsets across participants. The split-half reliability correlation coefficient was permuted 1,000 times and the average of the correlations is reported for each independent measure. Before averaging, all correlations were Fisher Z-transformed and then transformed back after averaging. In order to correct for underestimation resulting from splitting the number of observations in half, the Spearman-Brown correction was applied (Parsons et al., 2019). Reliability was not computed for sEBR during the rest period since it contained no trials. Descriptive statistics and split-half reliability measures are summarized in **Tables 1** and **2** for experiment 1 and experiment 2, respectively. Correlations between sEBR and performance with associated values of p and confidence intervals are summarized in **Tables 3** and **4** for experiment 1 and experiment 2, respectively. Post-hoc statistical power calculations using as the effect size the correlation between sEBR during the Delay period and task accuracy showed inadequate power for experiment 1 ($N = 16$, 1-β error probability = 0.53, critical $r = 0.49$). Power for experiment 2 was good ($N = 34$, 1-β error probability = 0.87, critical $r = 0.33$). Given that experiments 1 and 2 utilized different methods of quantifying blinks (camera based vs. EOG) but a similar task design, the two sample sizes were combined with very good power obtained ($N = 50$, 1-β error probability = 0.96, critical $r = 0.27$).

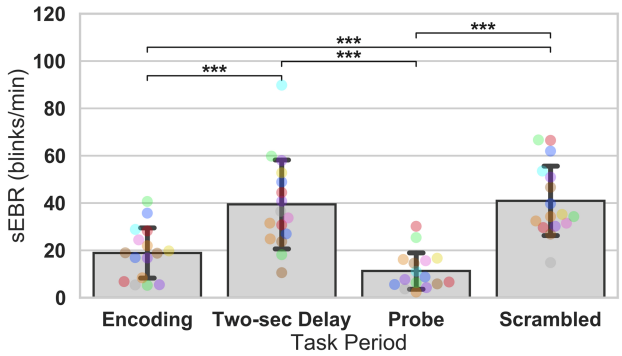


FIGURE 3 | ANOVA test of sEBR across task periods in Experiment 1. Bar plots show task period on the x-axis and sEBR (blinks/min) on the y-axis. Delay period sEBR was significantly greater than Encoding and Probe sEBR. Scrambled sEBR was also significantly greater than Encoding and Probe sEBR. Error bars depict 95% confidence interval. Each colored circle represents an individual participant; some colors may be presented twice in one bar due to limited primary colors available for display. Values of p were adjusted for comparing a family of 4. ***$p < 0.001$.

DISCUSSION

In the present study, we investigated the temporal fluctuations in sEBR across a WM paradigm and its relation to WM task accuracy in two experiments, inside and outside an MRI scanner, and using two methods of collecting sEBR. Using the same Sternberg working memory paradigm, we observed a strong positive relationship between sEBR and task performance only

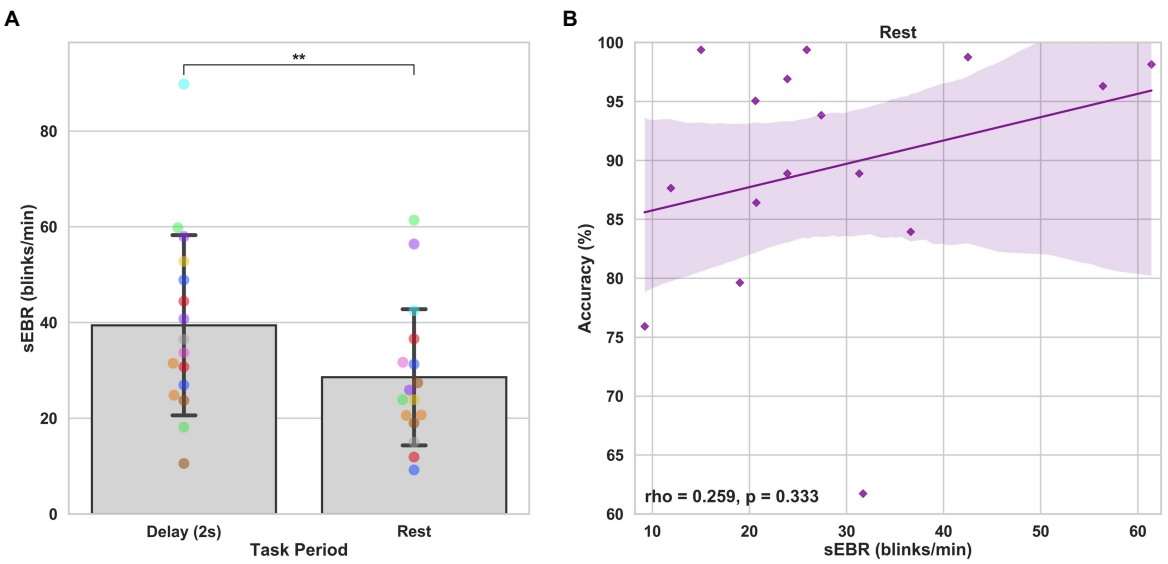


FIGURE 4 | Paired T tests between Delay period sEBR and Rest sEBR and correlation between Rest sEBR and task accuracy in Experiment 1. **(A)** Bar plots show task period on the x-axis and sEBR on the y-axis. Delay period sEBR was significantly higher than Rest sEBR. Error bars depict 95% confidence interval. **(B)** Correlation plot of sEBR during the Rest period on the x-axis and task accuracy on the y-axis. Fitted line represents linear regression model fit. Shaded region depicts 95% confidence interval. **$p < 0.01$.

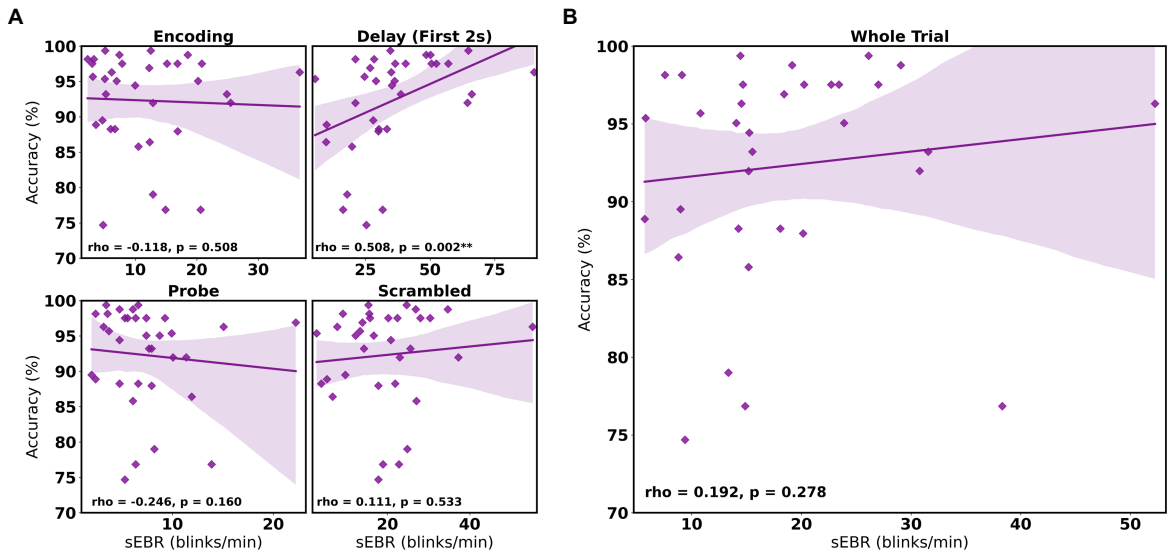


FIGURE 5 | Correlation between sEBR during different phases of the task and task accuracy in Experiment 2. Correlation plots show sEBR on the x-axis and task accuracy on the y-axis. **(A)** These four plots are encoding (top left), the first 2 s of the delay (top right), probe (bottom left), and scrambled (bottom right) periods. The delay period shows a strong positive correlation ($p = 0.002$, significant after a multiple comparison correction) between task accuracy and sEBR during the first 2 s of the delay period. **(B)** This plot represents the relationship between sEBR during the whole trial and task accuracy. Fitted line represents linear regression model fit. Shaded region depicts 95% confidence interval. Values of p for **(A)** after Bonferroni correction: $^{**}p < 0.0025$.

TABLE 2 | Descriptive statistics and split-half reliability for Experiment 2.

Variable	*n*	*M*	*SD*	Skewness	Kurtosis	Split-half coefficient
Task Accuracy (%)	34	92.30	6.956	−1.266	0.765	0.678
Whole Trial sEBR	34	18.486	9.843	1.453	3.023	0.992
Encoding sEBR	34	11.63	7.979	1.153	1.433	0.971
Delay sEBR	34	35.605	18.336	0.918	1.127	0.987
Probe sEBR	34	7.298	4.112	1.67	4.208	0.992
Scrambled sEBR	34	19.474	10.582	1.092	2.72	0.992
Rest sEBR	34	17.721	11.121	0.911	0.256	

M and SD represent mean and standard deviation, respectively. Split-half reliability based on 1,000 bootstrap replicates.

during the WM task delay in both experiments. We also found a significant difference in sEBR between task phases and a difference between Delay period sEBR and baseline sEBR.

Our first hypothesis was that phasic sEBR during the Delay period of the WM task would be positively correlated with task accuracy and that we would also observe a non-linear relationship where high and low sEBR would be predictive of low performance. We observed a strong positive correlation between sEBR during the Delay period in both Experiment 1 and Experiment 2 with task accuracy. However, only in Experiment 2 was this relationship significant. We believe that the lack of significance in Experiment 1 is due to the smaller sample size and thus lack of power, which our formal post-hoc power analyses confirmed. While the sample size in Experiment 2 was also small, we observed a similar correlation and reliability statistic as well as higher power while recording sEBR using a different method (electrooculogram instead of camera-based eye-tracking hardware). Nevertheless, a correlation between sEBR and task performance of approximately 0.50 is a very high correlation in psychology, especially between a behavioral measure and a physiological measure (Gignac and Szodorai, 2016). The replication of a similar correlation between Delay period sEBR and WM performance across two separate experiments strengthens our findings but with the confidence interval being so wide with the relatively small sample size one cannot be certain the true correlation is so large. Previous research has found that correlations begin to stabilize at even larger sample sizes (Schönbrodt and Perugini, 2013). Thus, future studies should include an additional experiment with high statistical power to replicate these findings and to determine whether the observed effect stabilizes with even larger sample sizes.

If we interpret sEBR as an indirect measure of striatal DA activity, as other studies have postulated, we can speculate that higher sEBR during the WM delay was correlated with task accuracy due to DA regulating the maintenance and

updating of representations in WM (Westbrook and Braver, 2016). The other results support this idea since no other task period was significantly correlated with task accuracy. Many studies that have investigated the relationship between sEBR and cognitive functions have used baseline levels of sEBR taken before or after tasks in their analysis (Tharp and Pickering, 2011; Zhang et al., 2015; Unsworth et al., 2019b). However, we show that while the WM task Delay period sEBR was correlated positively with task accuracy, baseline levels of sEBR were not. Our results highlight the importance of examining phasic and tonic sEBR when investigating the relationships between sEBR and other cognitive functions. The results also highlight that blinking may be an important component of working memory function; however, future studies, including within-subject analyses using larger number of trials, are needed to understand the role of blinking during WM maintenance. Additionally, future studies should investigate whether higher blink rates during the WM delay lead to a correct response. Since task difficulty was not controlled for in this study, participants' task performance in both experiments was relatively high (see **Tables 1** and **2**). These limitations make our current dataset incapable of investigating these questions.

We also investigated the proposed "Inverted U-shape" relationship between DA and WM performance by computing a polynomial regression model on sEBR during the delay and task accuracy (Cools and D'Esposito, 2011). Though the model showed a non-linear trend in Experiment 2, the model was not significant. We believe that failure to achieve non-linear model significance was due to lack of extreme (sub- and supraoptimal) sEBRs observed in the pool of participants, which are typically found in clinical populations (e.g., with Schizophrenia; Adamson, 1995; Swarztrauber and Fujikawa, 1998). Future studies should investigate sEBR with healthy subjects and with subjects that have been observed to have extreme sEBR in order to have a wider variety of sEBRs and to better understand its connection with DA. Additionally, other methods of DA measures could be used to investigate DA during the delay period, such as correlations with neuromelanin-sensitive MRI, which can detect neuromelanin, a product of dopamine metabolism (Cassidy et al., 2019).

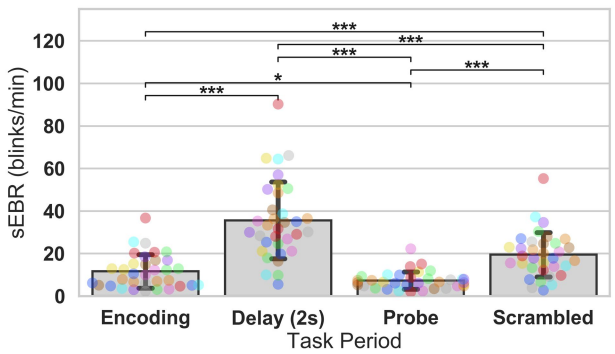


FIGURE 6 | ANOVA test of sEBR across task periods in Experiment 2. Bar plots show task period on the x-axis and sEBR on the y-axis. Delay period sEBR was significantly greater than Encoding, Probe, and Scrambled sEBR. Scrambled sEBR was also significantly greater than Encoding and Probe sEBR. Encoding sEBR was significantly greater than Probe sEBR. Error bars depict 95% confidence interval. Each colored circle represents an individual participant; some colors may be presented twice in one bar due to limited primary colors available for display. Values of p were adjusted for comparing a family of 4. $^{*}p < 0.05$, and $^{***}p < 0.001$.

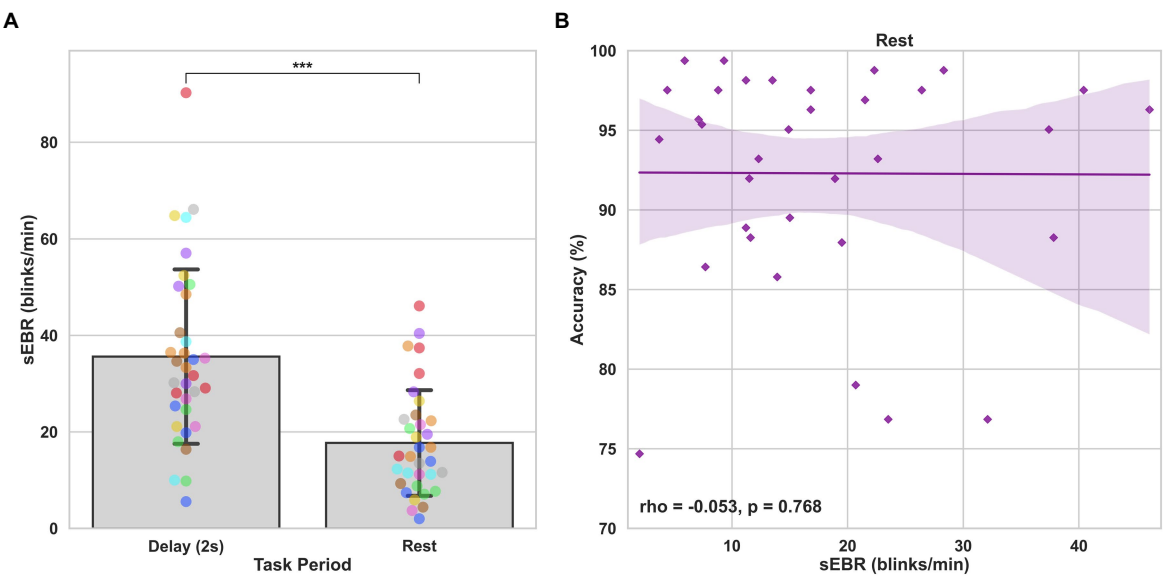


FIGURE 7 | Paired T tests between Delay period sEBR and Rest sEBR and correlation between Rest sEBR and task accuracy in Experiment 2. **(A)** Bar plots show task period on the x-axis and sEBR on the y-axis. Delay period sEBR was significantly higher than Rest sEBR. Error bars depict 95% confidence interval. **(B)** Correlation plot of sEBR during the Rest period on the x-axis and task accuracy on the y-axis. Fitted line represents linear regression model fit. Shaded region depicts 95% confidence interval. $^{***}p < 0.001$.

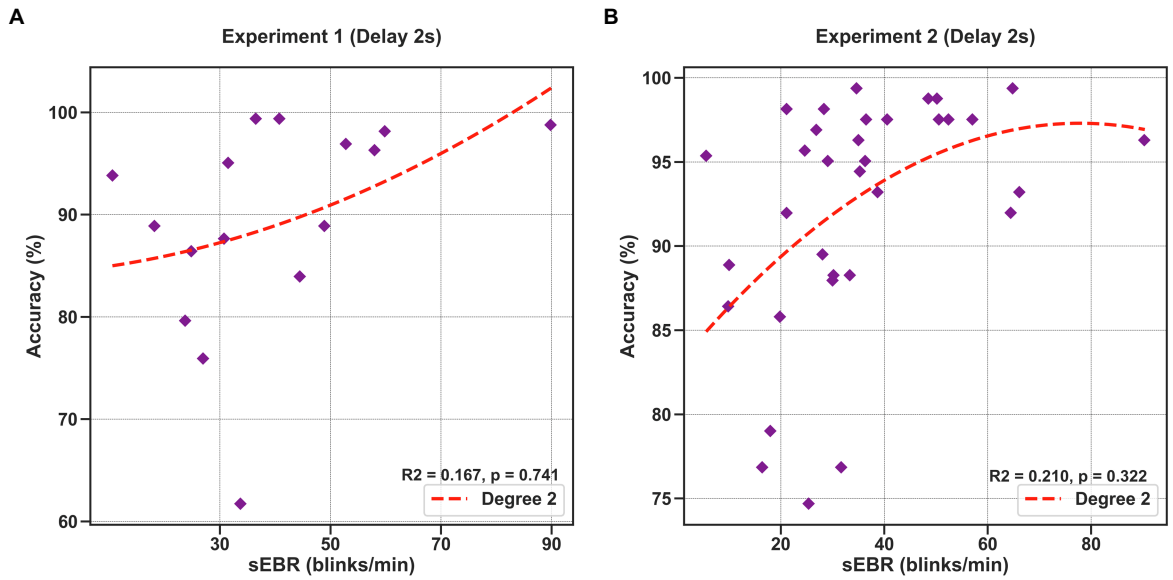


FIGURE 8 | Polynomial regression model between task accuracy and sEBR during the first 2 s of the delay for Experiment 1 and Experiment 2. Regression plots show sEBR during the first 2 s of the Delay on the x-axis and task accuracy on the y-axis. **(A)** Polynomial regression model fitted on sEBR during the Delay and task accuracy in Experiment 1. **(B)** Polynomial regression model fitted on sEBR during the Delay and task accuracy in Experiment 2. Fitted red line represents polynomial regression model fit. The relationship between sEBR and WM performance appears to be non-linear and explains about 20% of the variance in Experiment 2 but does not reach significance.

TABLE 3 | Correlations between sEBR and performance for Experiment 1.

Variable	Spearman's rho	Value of *p*	95% CI
Encoding sEBR	−0.078	0.774	[−0.569, 0.490]
Delay sEBR	0.526	0.036	[0.121, 0.783]
Probe sEBR	−0.187	0.488	[−0.699, 0.401]
Scrambled sEBR	0.054	0.843	[−0.477, 0.623]
Rest sEBR	0.259	0.333	[−0.285, 0.740]
Whole trial sEBR	0.149	0.582	[−0.369, 0.676]

CI, confidence intervals. CIs based on 1,000 bootstrap replicates.

TABLE 4 | Correlations between sEBR and performance for Experiment 2.

Variable	Spearman's rho	Value of *p*	Confidence Intervals
Encoding sEBR	−0.118	0.508	[−0.418, 0.224]
Delay sEBR	0.508	0.002	[0.213, 0.699]
Probe sEBR	−0.246	0.16	[−0.527, 0.082]
Scrambled sEBR	0.111	0.533	[−0.228, 0.423]
Rest sEBR	−0.053	0.768	[−0.400, 0.310]
Whole trial sEBR	0.192	0.278	[−0.148, 506]

CI, confidence intervals. CIs based on 1,000 bootstrap replicates.

Our second hypothesis was that we would see significant differences in sEBR across the WM task as well as between sEBR during Rest and during the Delay period. We found sEBR to be the lowest during periods like Encoding and Probe in both Experiments, while sEBR during the Delay was the highest. Our results support previous findings which found task-related modulation of sEBR (Siegle et al., 2008; Oh et al., 2012). Prior work has found sEBR to be lower during tasks that require visual attention (Fukuda et al., 2005; Oh et al., 2012). This would explain the lower sEBR's observed during the Encoding period when participants are encoding information into WM and during the Probe period where participants are retrieving information from WM. We also found that sEBR was the highest during the Delay period when participants were maintaining information in WM. This was also demonstrated in a different study which investigated sEBR during an A-not-B WM task where infants had to search for a hidden toy by making an eye movement to one of two locations (Bacher et al., 2017). Higher sEBR during the WM delay could be due to DA regulating the maintenance and updating of representations in WM (Westbrook and Braver, 2016), but this remains speculation until further studies directly measure dopaminergic activity during task performance. Our results further support this interpretation since Delay period sEBR was significantly higher than baseline sEBR during the Rest period. Lower sEBR during the Rest period could be explained since there is no need to update or maintain WM during this period.

To conclude, we investigated temporal changes of sEBR during different phases of a WM task and its relation to WM performance. We observed a significant positive correlation between sEBR and WM task performance during the Delay period, but not during the other phases of the task. Additionally, we found evidence for an association of sEBR during both stimulus encoding and WM probe retrieval, potential reflecting visual attention. To the best of our knowledge, this is the first study to investigate phasic and tonic sEBR during different phases of a WM task using complex visual scenes. Future studies

should continue to investigate sEBRs in relation to direct measures of cortical (especially PFC) and subcortical dopamine and assess linear and non-linear relationships to task performance in healthy and clinical populations (e.g., Schizophrenia and Parkinson's disease).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the City University of New York Human Research Protection Program (CUNY HRPP IRB). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TE designed the study. JO, CR, and BG performed the research. JO analyzed and interpreted the data, prepared the figures, and wrote the final manuscript. CR, BG, and TE edited and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R56MH116007 (TE).

REFERENCES

Adamson, T. A. (1995). Changes in blink rates of Nigerian schizophrenics treated with chlorpromazine. *West Afr. J. Med.* 14, 194–197.

Bacher, L. F., and Allen, K. J. (2009). Sensitivity of the rate of spontaneous eye blinking to type of stimuli in young infants. *Dev. Psychobiol.* 51, 186–197. doi: 10.1002/dev.20357

Bacher, L. F., Retz, S., Lindon, C., and Bell, M. A. (2017). Intraindividual and Interindividual differences in spontaneous eye blinking: relationships to working memory performance and frontal EEG asymmetry. *Infancy* 22, 150–170. doi: 10.1111/infa.12164

Baddeley, A. (1992). Working memory. *Science* 255, 556–559. doi: 10.1126/science.1736359

Barbato, G., Ficca, G., Muscettola, G., Fichele, M., Beatrice, M., and Rinaldi, F. (2000). Diurnal variation in spontaneous eye-blink rate. *Psychiatry Res.* 93, 145–151. doi: 10.1016/S0165-1781(00)00108-6

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., and Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Mov. Disord.* 12, 1028–1034. doi: 10.1002/mds.870120629

Cassidy, C. M., Zucca, F. A., Girgis, R. R., Baker, S. C., Weinstein, J. J., Sharp, M. E., et al. (2019). Neuromelanin-sensitive MRI as a noninvasive proxy measure of dopamine function in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 116, 5108–5117. doi: 10.1073/pnas.1807983116

Colzato, L. S., Van Den Wildenberg, P. M., Van Wouwe, N. C., Pannebakker, M. M., and Hommel, B. (2009). Dopamine and inhibitory action control: evidence from spontaneous eye blink rates. *Exp. Brain Res.* 196, 467–474. doi: 10.1007/s00221-009-1862-x

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X. L., Wang, M., et al. (2018). Persistent spiking activity underlies working memory. *J. Neurosci.* 38, 7020–7028. doi: 10.1523/JNEUROSCI.2486-17.2018

Cools, R., and D'esposito, M. (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biol. Psychiatry* 69, E113–E125. doi: 10.1016/j.biopsych.2011.03.028

Cools, R., Gibbs, S. E., Miyakawa, A., Jagust, W., and D'esposito, M. (2008). Working memory capacity predicts dopamine synthesis capacity in the human striatum. *J. Neurosci.* 28, 1208–1212. doi: 10.1523/JNEUROSCI.4475-07.2008

Cools, R., and Robbins, T. W. (2004). Chemistry of the adaptive mind. *Philos. Trans. A. Math. Phys. Eng. Sci.* 362, 2871–2888. doi: 10.1098/rsta.2004.1468

Courtney, S. M., Petit, L., Maisog, J. M., Ungerleider, L. G., and Haxby, J. V. (1998). An area specialized for spatial working memory in human frontal cortex. *Science* 279, 1347–1351. doi: 10.1126/science.279.5355.1347

Dang, L. C., Samanez-Larkin, G. R., Castrellon, J. J., Perkins, S. F., Cowan, R. L., Newhouse, P. A., et al. (2017). Spontaneous eye blink rate (EBR) is uncorrelated with dopamine D2 receptor availability and Unmodulated by dopamine Agonism in healthy adults. *eNeuro* 4:ENEURO.0211-17.2017. doi: 10.1523/ENEURO.0211-17.2017

Dauer, W., and Przedborski, S. (2003). Parkinson's disease: mechanisms and models. *Neuron* 39, 889–909. doi: 10.1016/S0896-6273(03)00568-3

De Frias, C. M., Marklund, P., Eriksson, E., Larsson, A., Oman, L., Annerbrink, K., et al. (2010). Influence of COMT gene polymorphism on fMRI-assessed sustained and transient activity during a working memory task. *J. Cogn. Neurosci.* 22, 1614–1622. doi: 10.1162/jocn.2009.21318

Doughty, M. J., and Naase, T. (2006). Further analysis of the human spontaneous eye blink rate by a cluster analysis-based approach to categorize individuals with 'normal'versus 'frequent'eye blink activity. *Eye Contact Lens* 32, 294–299. doi: 10.1097/01.icl.0000224359.32709.4d

Durstewitz, D., and Seamans, J. K. (2008). The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-O-Methyltransferase genotypes and schizophrenia. *Biol. Psychiatry* 64, 739–749. doi: 10.1016/j.biopsych.2008.05.015

Elsworth, J. D., Lawrence, M. S., Roth, R. H., Taylor, J. R., Mailman, R. B., Nichols, D. E., et al. (1991). D1 and D2 dopamine receptors independently regulate spontaneous blink rate in the vervet monkey. *J. Pharmacol. Exp. Ther.* 259, 595–600.

Frank, M. J., Loughry, B., and O'reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cogn. Affect. Behav. Neurosci.* 1, 137–160. doi: 10.3758/CABN.1.2.137

Fukuda, K., Stern, J. A., Brown, T. B., and Russo, M. B. (2005). Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. *Aviat. Space Environ. Med.* 76(Suppl.), C75–C85.

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* 61, 331–349. doi: 10.1152/jn.1989.61.2.331

Fuster, J. M., and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science* 173:652. doi: 10.1126/science.173.3997.652

Gignac, G. E., and Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personal. Individ. Differ.* 102, 74–78. doi: 10.1016/j.paid.2016.06.069

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Groman, S. M., James, A. S., Seu, E., Tran, S., Clark, T. A., Harpster, S. N., et al. (2014). In the blink of an eye: relating positive-feedback sensitivity to striatal dopamine D2-like receptors through blink rate. *J. Neurosci.* 34, 14443–14454. doi: 10.1523/JNEUROSCI.3037-14.2014

Hall, A. (1945). The origin and purposes of blinking. *Br. J. Ophthalmol.* 29, 445–467. doi: 10.1136/bjo.29.9.445

Hazy, T. E., Frank, M. J., and O'reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience* 139, 105–118. doi: 10.1016/j.neuroscience.2005.04.067

Howes, O., Mccutcheon, R., and Stone, J. (2015). Glutamate and dopamine in schizophrenia: An update for the 21st century. *J. Psychopharmacol.* 29, 97–115. doi: 10.1177/0269881114563634

Jongkees, B. J., and Colzato, L. S. (2016). Spontaneous eye blink rate as predictor of dopamine-related cognitive function-A review. *Neurosci. Biobehav. Rev.* 71, 58–82. doi: 10.1016/j.neubiorev.2016.08.020

Jutkiewicz, E. M., and Bergman, J. (2004). Effects of dopamine D1 ligands on eye blinking in monkeys: efficacy, antagonism, and D1/D2 interactions. *J. Pharmacol. Exp. Ther.* 311, 1008–1015. doi: 10.1124/jpet.104.071092

Kimberg, D. Y., D'esposito, M., and Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport* 8, 3581–3585. doi: 10.1097/00001756-199711100-00032

Landau, S. M., Lal, R., O'Neil, J. P., Baker, S., and Jagust, W. J. (2009). Striatal dopamine and working memory. *Cereb. Cortex* 19, 445–454. doi: 10.1093/cercor/bhn095

Maffei, A., and Angrilli, A. (2018). Spontaneous eye blink rate: An index of dopaminergic component of sustained attention and fatigue. *Int. J. Psychophysiol.* 123, 58–63. doi: 10.1016/j.ijpsycho.2017.11.009

Norn, M. S. (1969). Desiccation of the precorneal film. I. Corneal wetting-time. *Acta. Ophthalmol.* 47, 865–880. doi: 10.1111/j.1755-3768.1969.tb03711.x

Oh, J., Jeong, S. Y., and Jeong, J. (2012). The timing and temporal patterns of eye blinking are dynamically modulated by attention. *Hum. Mov. Sci.* 31, 1353–1365. doi: 10.1016/j.humov.2012.06.003

Parsons, S., Kruijt, A.-W., and Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Adv. Methods Pract. Psychol. Sci.* 2, 378–395. doi: 10.1177/2515245919879695

Rac-Lubashevsky, R., Slagter, H. A., and Kessler, Y. (2017). Tracking real-time changes in working memory updating and gating with the event-based eye-blink rate. *Sci. Rep.* 7:2547. doi: 10.1038/s41598-017-02942-3

Reddy, V. C., Patel, S. V., Hodge, D. O., and Leavitt, J. A. (2013). Corneal sensitivity, blink rate, and corneal nerve density in progressive Supranuclear palsy and Parkinson disease. *Cornea* 32, 631–635. doi: 10.1097/ICO.0b013e3182574ade

Roberts, A. C., Robbins, T. W., and Weiskrantz, L. E. (1998). *The Prefrontal Cortex: Executive and Cognitive Functions*. United Kingdom: Oxford University Press.

Sawaguchi, T. (2001). The effects of dopamine and its antagonists on directional delay-period activity of prefrontal neurons in monkeys during an oculomotor delayed-response task. *Neurosci. Res.* 41, 115–128. doi: 10.1016/S0168-0102(01)00270-X

Schönbrodt, F. D., and Perugini, M. (2013). At what sample size do correlations stabilize? *J. Res. Pers.* 47, 609–612. doi: 10.1016/j.jrp.2013.05.009

Seamans, J. K., and Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog. Neurobiol.* 74, 1–58. doi: 10.1016/j.pneurobio.2004.05.006

Sescousse, G., Ligneul, R., Van Holst, R. J., Janssen, L. K., De Boer, F., Janssen, M., et al. (2018). Spontaneous eye blink rate and dopamine synthesis capacity: preliminary evidence for an absence of positive correlation. *Eur. J. Neurosci.* 47, 1081–1086. doi: 10.1111/ejn.13895

Siegle, G. J., Ichikawa, N., and Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology* 45, 679–687. doi: 10.1111/j.1469-8986.2008.00681.x

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153:652. doi: 10.1126/science.153.3736.652

Stewart, C. V., and Plenz, D. (2006). Inverted-U profile of dopamine-NMDA-mediated spontaneous avalanche recurrence in superficial layers of rat prefrontal cortex. *J. Neurosci.* 26, 8148–8159. doi: 10.1523/JNEUROSCI.0723-06.2006

Swarztrauber, K., and Fujikawa, D. G. (1998). An electroencephalographic study comparing maximum blink rates in schizophrenic and nonschizophrenic psychiatric patients and nonpsychiatric control subjects. *Biol. Psychiatry* 43, 282–287. doi: 10.1016/S0006-3223(97)00028-0

Taylor, J. R., Elsworth, J. D., Lawrence, M. S., Sladek, J. R., Roth, R. H., and Redmond, D. E. (1999). Spontaneous blink rates correlate with dopamine levels in the caudate nucleus of MPTP-treated monkeys. *Exp. Neurol.* 158, 214–220. doi: 10.1006/exnr.1999.7093

Tharp, I. J., and Pickering, A. D. (2011). Individual differences in cognitive-flexibility: the influence of spontaneous eyeblink rate, trait psychoticism and working memory on attentional set-shifting. *Brain Cogn.* 75, 119–125. doi: 10.1016/j.bandc.2010.10.010

Tsubota, K., Hata, S., Okusawa, Y., Egami, F., Ohtsuki, T., and Nakamori, K. (1996). Quantitative videographic analysis of blinking in normal subjects and patients with dry eye. *Arch. Ophthalmol.* 114, 715–720. doi: 10.1001/archopht.1996.01100130707012

Unsworth, N., Miller, A. L., and Robison, M. K. (2019a). Individual differences in encoding strategies and free recall dynamics. *Q. J. Exp. Psychol.* 72, 2495–2508. doi: 10.1177/1747021819847441

Unsworth, N., Robison, M. K., and Miller, A. L. (2019b). Individual differences in baseline oculometrics: examining variation in baseline pupil diameter, spontaneous eye blink rate, and fixation stability. *Cogn. Affect. Behav. Neurosci.* 19, 1074–1093. doi: 10.3758/s13415-019-00709-z

Van Slooten, J. C., Jahfari, S., and Theeuwes, J. (2019). Spontaneous eye blink rate predicts individual differences in exploration and exploitation during reinforcement learning. *Sci. Rep.* 9:17436. doi: 10.1038/s41598-019-53805-y

Wager, T. D., and Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis. *Cogn. Affect. Behav. Neurosci.* 3, 255–274. doi: 10.3758/CABN.3.4.255

Westbrook, A., and Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron* 89, 695–710. doi: 10.1016/j.neuron.2015.12.029

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). "Sun database: Large-scale scene recognition from abbey to zoo", in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: IEEE*; June 13, 2010, 3485–3492.

Zhang, T., Mou, D., Wang, C., Tan, F., Jiang, Y., Lijun, Z., et al. (2015). Dopamine and executive function: increased spontaneous eye blink rates correlate with better set-shifting and inhibition, but poorer updating. *Int. J. Psychophysiol.* 96, 155–161. doi: 10.1016/j.ijpsycho.2015.04.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership