Heliyon 8 (2022) e09207

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Applying machine learning to dissociate between stroke patients and healthy controls using eye movement features obtained from a virtual reality task

Veerle H.E.W. Brouwer [a], Sjoerd Stuit [a], Alex Hoogerbrugge [a], Antonia F. Ten Brink [a], Isabel K. Gosselt [b], Stefan Van der Stigchel [a], Tanja C.W. Nijboer [a,b,c,*]

[a] Department of Experimental Psychology, Helmholtz Institute, Utrecht University, Heidelberglaan 1, 3584 CS, Utrecht, Netherlands
[b] Center of Excellence for Rehabilitation Medicine, UMC Utrecht Brain Center, University Medical Center Utrecht, De Hoogstraat Rehabilitation, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands
[c] Department of Rehabilitation, Physical Therapy Science & Sports, UMC Utrecht Brain Center, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands

ARTICLE INFO

Keywords:
Stroke
Cognitive assessment
Virtual reality
Eye tracking
Machine learning

ABSTRACT

Conventional neuropsychological tests do not represent the complex and dynamic situations encountered in daily life. Immersive virtual reality simulations can be used to simulate dynamic and interactive situations in a controlled setting. Adding eye tracking to such simulations may provide highly detailed outcome measures, and has great potential for neuropsychological assessment. Here, participants (83 stroke patients and 103 healthy controls) we instructed to find either 3 or 7 items from a shopping list in a virtual super market environment while eye movements were being recorded. Using Logistic Regression and Support Vector Machine models, we aimed to predict the task of the participant and whether they belonged to the stroke or the control group. With a limited number of eye movement features, our models achieved an average Area Under the Curve (AUC) of .76 in predicting whether each participant was assigned a short or long shopping list (3 or 7 items). Identifying participant as either stroke patients and controls led to an AUC of .64. In both classification tasks, the frequency with which aisles were revisited was the most dissociating feature. As such, eye movement data obtained from a virtual reality simulation contain a rich set of signatures for detecting cognitive deficits, opening the door to potential clinical applications.

1. Introduction

One of the consequences of stroke is cognitive impairment, which interferes with a wide range of complex activities in daily life, such as work or school, social events, and travelling. Cognitive functioning is usually assessed by means of a neuropsychological assessment. One of the drawbacks of a conventional neuropsychological assessment, however, is that it is notoriously different from the complex, dynamic situations of daily living. Another drawback is the relative crudeness of the outcome measures; for most neuropsychological tests, accuracy and/or total duration of the task constitute the only outcome measures to be compared to the range of scores that are considered 'normal'.

There is an urgent need for better, easy to use, and ecologically valid methods to assess cognitive skills in more complex environments. Recently, advancements in virtual reality (VR) have resulted in new assessment methods with high ecological validity, holding promising opportunities for developing accurate assessments of cognitive functions (Parsons, 2015; Rizzo et al., 2004). VR simulations can mimic daily life environments that are immersive, and dynamic while also being highly controllable (Brooks and Rose, 2003; Carassa et al., 2005; You et al., 2005). Implementing eye tracking in VR simulations has an even greater potential for the assessment of cognitive functioning (Clay et al., 2019). Because of the strong overlap between the oculomotor system and various cognitive domains, such as visual attention (Corbetta et al., 1998; Rizzolatti et al., 1987), executive functioning (Bosch et al., 2013; Everling and Fischer, 1998), and (working) memory (Richardson and Spivey, 2000; Van der Stigchel and Hollingworth, 2018), eye tracking is the ideal technique to index these cognitive processes in both healthy and clinical

* Corresponding author.
E-mail address: t.c.w.nijboer@uu.nl (T.C.W. Nijboer).

https://doi.org/10.1016/j.heliyon.2022.e09207
Received 24 March 2021; Received in revised form 27 May 2021; Accepted 24 March 2022

populations. For example, when searching a cluttered scene, the current search goal and search strategy of an observer are directly reflected in oculomotor features, such as saccade amplitude and the inter-saccadic interval (Mills et al., 2011). Also, refixations have been associated with inefficient search behaviour, as it is generally assumed that a refixation is performed when not all required information at the re-fixated location has been internalised during the previous fixation, which might be related to deficits in either encoding or maintaining recently acquired information in working memory (Najemnik and Geisler, 2005).

To this end, gaze behaviour allows for the extraction of various features which can be used to examine the mechanisms underlying behaviour in a natural environment, such as the number of refixations, fixation duration, and pupil size, which are potential novel outcome measures for cognitive functions. In the current study, eye tracking was integrated in a VR simulation and we aim to determine which eye movement features could best dissociate between different task requirements and best categorise stroke patients and healthy control participants, using machine learning (see also Lagun et al., 2011, patients with mild cognitive impairment; Charron et al., 2010; Cyr et al., 2009; Delazer et al., 2018; Husain et al., 2001; Walle et al., 2019, stroke patients). Machine learning approaches, in comparison to more conventional statistical approaches, will not only classify cases, they can also be applied when data is normally distributed and outliers are relatively common, as can be the case with patient data. These methods, using eye movement measures, have been previously used to differentiate between various neurodegenerative and neurodevelopmental disorders (Carette et al., 2019; Lagun et al., 2011; Pusiol et al., 2016) and to decode the task of the observer (Borji et al., 2015; Borji and Itti, 2014; Kootstra et al., 2020). Given the recent rise of affordable and accessible eye tracking hardware, there is great promise in applying machine learning techniques to improve neuropsychological assessment using non-invasive, sensitive measures.

2. Materials and methods

2.1. Participants

Stroke patients admitted for inpatient or (former) outpatient rehabilitation were asked to participate in this study. Outpatients and former outpatients were recruited at the University Medical Centre Utrecht (UMC Utrecht) from June 2016 to September 2018 and at the Hoogstraat Rehabilitation Centre from May 2018 to November 2018. Inpatients were recruited at the Hoogstraat Rehabilitation Centre from February 2018 to July 2019. Patients were included according to the following criteria: (1) clinically diagnosed stroke, first or recurrent, confirmed by an MRI or CT scan; (2) $\geq$ 18 years of age; and (3) being mentally and physically able to participate (evaluated by a neuropsychologist and rehabilitation physician). For patients who finished their rehabilitation programme, additional inclusion criteria were (4) having no planned/completed neuropsychological assessment for clinical purposes in the coming or past 3 months; and (5) having subjective cognitive complaints in daily life. Exclusion criteria for all patients were: (1) diagnosis of epilepsy; (2) diagnosis of visuospatial neglect; (3) motor problems that prevent the use of a controller; (4) comprehension and communication problems that prevent the task from being properly understood and executed; and (5) severe problems in arousal, alertness and/or vigilance that prevent patients from performing the task adequately (e.g., falling asleep during assessment). Inpatients and (former) outpatients recruited in the UMC Utrecht were assessed by a rehabilitation physician with respect to the in- and exclusion criteria. Outpatients at the Hoogstraat Rehabilitation Centre were screened on the inclusion and exclusion criteria by reviewing the medical records.

Neurologically healthy controls were recruited among relatives of the staff and via social media. Healthy controls were included based on the following criteria: (1) $\geq$ 18 years of age; (2) no history of neurological and/or psychiatric disorders; and (3) no physical conditions interfering with the testing procedure.

All participants gave written informed consent. The experiment was performed in accordance with the Declaration of Helsinki. The Medical Research Ethics Committee of the UMC Utrecht approved the research protocols (number 15-761/C, 17-667, 18-210, 18-799, and 19-112).

2.2. Procedure and medical characteristics

For (former) outpatients, the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) was administered before starting the VR experiment. For inpatients, the most recent recorded MoCA scores were retrieved from the medical record. Sex, age, and level of education were collected before the VR experiment by a short questionnaire. A Dutch classification system consisting of seven levels was used to assess the level of education (Verhage, 1965). These levels were converted into three categories: low (Verhage 1-4), average (Verhage 5), and high (Verhage 6-7). For stroke patients, the following information was obtained from the medical records: days post-stroke onset, lesion side, stroke type, and stroke history. The entire experiment took approximately 1 h.

2.3. Virtual reality simulation: virtual supermarket

A virtual supermarket was developed by Atoms2Bits for commercial purposes, using the Unity game development platform. It was adapted for research and potential clinical purposes in close collaboration with Utrecht University, UMC Utrecht, and De Hoogstraat Rehabilitation Centre. The virtual environment was based on a real-life Dutch supermarket. The total floor space was 49.75 by 29.75 m with a height of 4 m. The supermarket spanned eleven shopping aisles, seven of which had shelves on either side. The other aisles only had shelves one side, thus the supermarket contained eighteen groups of shelves, totaling 12,000 selectable products, sorted in different sections (e.g., bakery and produce section) and 8 cash registers.

The VR environment was run on an Intel® Core™ i7-4790K CPU with 16GB of RAM memory and a Geforce® GTX 970 video card. The operating system was Windows 10 Pro, 64-bit. Participants were provided with an HTC-Vive head-mounted display, which displayed at a resolution of 1080 $\times$ 1200 pixels per eye with a 110° diagonal field of view and a refresh rate of approximately 90 Hz. The participant interacted with the virtual environment by using two HTC Vive controllers, each containing buttons and a trigger. The participant could freely move through the supermarket, with a maximum speed of 1.8 km/h, by pulling the trigger of the controller. A laser appeared by pushing a button on the controller. Participants were instructed to point the laser at a product, fixate their gaze at the same spot, and hold the button until the controller briefly vibrated – indicating that the product was registered successfully. All participants performed the task while sitting on a desk chair. To track the precise location and orientation of the head-mounted display and the controllers, two base stations were installed in the upper corners of the playing field, creating a virtual space of 360°, run on SteamVR software.

Eye movements were recorded using HTC Vive binocular add-on eye trackers provided by Pupil Labs (Pupil Labs, 2020). The eye trackers included binocular 200 Hz eye tracking cameras with infrared illuminators for each eye. The eye trackers were connected via USB cable to the computer, and data was configured using Pupil software version 0.9.14.7, developed by Pupil Labs (Kassner et al., 2014). The software also created a log with timestamps of when the experiment started and ended and when products were selected with the laser.

2.4. Task

Participants were familiarised with a shopping list that included both verbal labels and images, which they had to learn by heart. The list consisted either of 3 products (short shopping list) or 7 products (long shopping list). This list was presented three times and the experimenter named the products out loud.

Before entering the VR supermarket, participants had the opportunity to get familiar with the VR set up (head-mounted display, controllers) by moving around in a generic virtual environment. Next, a calibration and validation procedure was started to calibrate the eye trackers, which was repeated if validation was insufficient. Validation was performed by the experimenter, who instructed the participant to point the laser at each of the calibration locations and to fixate at the same point. The experimenter then visually assessed whether the gaze location as shown in real-time by the eye tracking software was indeed located within a small area around the intended location. After calibration of the eye trackers, and after participants indicated being confident enough to start the task, the experimenter would let them enter the VR supermarket. Participants were told that the assessment would start as soon as they passed the entry gates and that they had to pass the cash registers to finish the task. The task with the short list was part of a feasibility study, and as such participants were allowed to search for 15 min. In the task with the long list, participants were allowed to search for 7 min. After this maximum duration, participants were instructed to go the check-out, adding another couple of minutes to the total duration spent in the virtual environment.

2.5. Data preprocessing and eye tracking features

For completed tasks, raw data files usually consisted of at least 50,000 rows of data points, each row containing characteristics such as the timestamp, navigational data (e.g., the foot- and head position) and the area in which the participant was located (supermarket, cash register, or calibration area), pupil confidence (indicator of the eye tracker's confidence that it is correctly measuring the pupil), and pupil loss (indicator whether a pupil is recorded or not). Furthermore, the data recorded by the eye trackers consisted of both 2D and 3D pupil position coordinates for the left and right eye. The 3D coordinates were based on the 3D gaze vector plotted out of the 2D pupil position, combined with the position of a participant's location in the supermarket. This gaze vector was intersected with objects in the 3D supermarket environment and was defined by the dimensions X (depth), Y (height), and Z (width). An additional column indicated the specific object at which a participant was fixating by relating the 3D gaze coordinates to items' positions.

To determine whether a fragment of eye tracking data was recorded within an actual shopping aisle, navigational data was analysed, which indicated at which specific timestamp the participant entered and left a shopping aisle. These timestamps were used to segment the eye tracking data, since the analysis of eye characteristics was only performed over data in which a participant could view products inside an aisle area.

A filtering procedure was executed to ensure sufficient monocular eye tracking quality. This procedure filtered all data points according to the following steps: firstly, all data points with a pupil confidence below .60 were removed (Pupil Labs, 2020; although the basis for this value remains unclear). The pupil confidence score ranges from 0 (lowest confidence) to 1 (highest confidence). Secondly, because in the current study fixations were extracted from one eye (i.e., monocular), all data points where pupil loss was marked as 'both' were removed, so that at least one eye was tracked at any given data point. Additionally, missing data points were removed if they were not already signified by pupil confidence $<.60$. Finally, for each dataset, the best eye for eye tracking was chosen, based on the highest remaining data count after the three filter steps.

Due to high variability in sampling rate and data loss, a custom dispersion-based algorithm was designed to extract fixations from the raw data. Fixations were calculated over the 3D gaze coordinates, which were intersected with objects in the virtual environment, and were defined such that the gaze may move less than 2 degrees of visual angle over a span of 100 ms to 4 s. This upper time limit was selected since the process of pointing the laser at an object until it vibrated could sometimes take several seconds. Additionally, refixations were defined as a fixation at a location which has been fixated before, within a margin of 0.05 coordinates (5 cm) around the original fixation location and no minimal time between fixation and refixation (although at least one other fixation had to occur in-between).

Four 'local' features were computed as a low-to intermediate-level of behaviour within each aisle visit: (1) the number of fixations per second, (2) the number of refixations per second, (3) the median fixation duration (in milliseconds), and (4) the median time between each fixation and its accompanying refixation (in seconds, hereafter termed *median time until refixations*). One 'global' feature was computed as a measure of intermediate-to high-level behaviour: (5) the mean number of times each aisle was revisited by the participant (i.e., thereby controlling for the total number of visited aisles).

2.6. Analyses

2.6.1. Demographics

Non-parametric tests (Mann-Whitney U for age and Chi-square test for sex and level of education) were used to compare demographic characteristics between stroke patients and healthy controls. For all tests, α was set to 0.05.

2.6.2. Machine learning

In order to determine whether categories (i.e., short-versus long list; stroke patients versus healthy controls) could be adequately identified based on eye tracking data, two commonly used classification algorithms were implemented: (1) a Logistic Regression model and (2) a Support Vector Machine. The reason for choosing two instead of one classification algorithm was motivated by the fact that it allows for an extra measure of validation for how well the data generalises to different models since we use different techniques to classify the data. Additionally, a Logistic Regression allows for intuitive insight into the contribution of each feature to the model's classification strategy. The algorithms were implemented using the open-source Scikit-learn software version 0.23.2 (Pedregosa et al., 2011) in Python version 3.8.

2.6.3. Data aggregation

Since each participant visited multiple aisles – but not always the same between participants – and the algorithm required one set of features per participant, each of the four local features needed to be aggregated across aisles within participants. Note that, henceforth, the five features as described in *section 2.5* will be termed 'base features' and the aggregated local features will be called 'intermediate features'. It is important to note that simply averaging a local feature over all aisles is not a sufficient descriptor, as two participants can have the same average number of fixations per second across aisles, but a very different standard deviation. To this end, fifteen statistical descriptors were selected to construct each intermediate feature, which attempted to describe the distribution of an aggregated local feature on as many aspects as possible: mean, median, variance, standard deviation, skew, kurtosis, range, 10^{th} percentile, 90^{th} percentile, interquartile range, mean absolute deviation, energy, entropy, uniformity and root mean square.

However, four local features described by fifteen statistical descriptors led to a set of 60 intermediate features. Since one of our objectives was to uncover which of the set of five base features contributed most to accurate classification, it was important to find a workable subset of useful model features. To this end, Principal Component Analysis (PCA) was implemented prior to each model execution, which reduced each local feature's fifteen statistical descriptors to three new components (termed C1, C2 and C3), leading to a total of thirteen features: twelve local features (via PCA) and one global feature (the mean number of times each aisle was revisited).

2.6.4. Model execution

All models were trained using a 70/30 stratified training-test set split, meaning that 70% of data was used to train the model and 30% was used to validate the model's accuracy after training. This split was stratified, which means that both the train- and test set maintained a roughly equal

proportion of data of both classes (long vs short list, patients vs controls). To tackle residual imbalances in classes, model accuracies were measured by calculating Area Under the Curve (AUC).

Since the data set was relatively small for machine-learning standards, and a model's performance can depend on its random initialisation, 50 independent iterations (runs) were performed for each model. In each run, (1) the train-test split was resampled, (2) PCA was fitted to the training set and applied to the test set, and (3) the model was initialized, trained, and tested. The models' average AUC and standard deviations are reported over these 50 runs.

To estimate the importance of each feature during classification, we used the absolute values of the Logistic Regression's coefficients of the 50 independent runs. Note that only coefficients from run's where the AUC was sufficiently above chance level were included (chance level = .50, threshold for inclusion >.65) since the coefficients from non-significant models are not interpretable. For the sake of interpretability, components of the same base feature are recombined when reporting feature importances. In order to avoid including features in our analyses which are very similar to each other, we report on the correlations between all five base features, regardless of group or task type.

3. Results

3.1. Participants

Figure 1 shows a flowchart of the data selection process. In total, 134 stroke patients participated in the study, and data from 83 stroke patients could be used for the analyses. Of the stroke patients, 62 did the task with the short list (3 products), and 21 did the task with the long list (7 products). Additionally, 123 healthy controls participated and data of 103 healthy controls were used for analyses. Of the healthy controls, 27 did the task with the short list, and 76 did the task with the long list.

In Table 1, the demographic and medical characteristics of the stroke patients and healthy controls are given. The stroke patients group had a lower level of education (χ^2 (2) = 28.3, $p < .001$) and were overall older ($U = 6030.5$, $p < .001$, $f = .70$) compared to healthy controls. There was no significant difference in distribution of sex between the groups ($p = .12$). The whole-task descriptive can be found in Table 2. Based on the significant differences between stroke patients and controls in terms of level of education and age, we tested whether the variables were related to the feature used the modelling approaches. The mean (*SD*) value of each feature over all participants, after aggregation by aisle and averaging over all aisles per participant can be found in Table 3. Only refixations per second and level of education had a significant relationship (Rho = 0.1571, p = 0.15) prior to correcting for multiple comparisons.

Table 1. Demographic and medical characteristics of the stroke patients and healthy controls. Means (*SD*) and percentages are provided.

	Stroke patients ($n = 83$)		Healthy controls ($n = 103$)	
		n		*n*
Age in years, mean (*SD*)	55.08 (*12.97*)	83	44.32 (*14.90*)	103
Sex (% male)	58	83	45	103
Level of education (%)		83		103
Low	15.7		1.0	
Average	26.5		8.7	
High	57.8		90.3	
Task with 3 products (%)	74.7	62	26.2	27
Task with 7 products (%)	25.3	21	73.8	76
Days post-stroke onset, mean (*SD*)	530.58 (*988.18*)	80	-	
Lesion side (%)		83		
Left	43.4		-	
Right	38.6		-	
Both	8.4		-	
Unknown	9.6		-	
Stroke type (%)		83		
Ischemic	59.0		-	
Haemorrhage	19.3		-	
Subarachnoid Haemorrhage	14.5		-	
Subdural Hematoma	2.4		-	
Sinus Thrombosis	1.2		-	
Unknown	3.6			
Stroke history (%)		83		
First	86.7			
Recurrent	9.6			
Unknown	3.6			
MoCA, mean (*SD*) (0-30)	24.4 (*4.1*)	79	28.3 (*2.0*)	10
Raw data loss, mean (%)	46.3	83	44.1	103

Notes. MoCA = Montreal Cognitive Assessment.

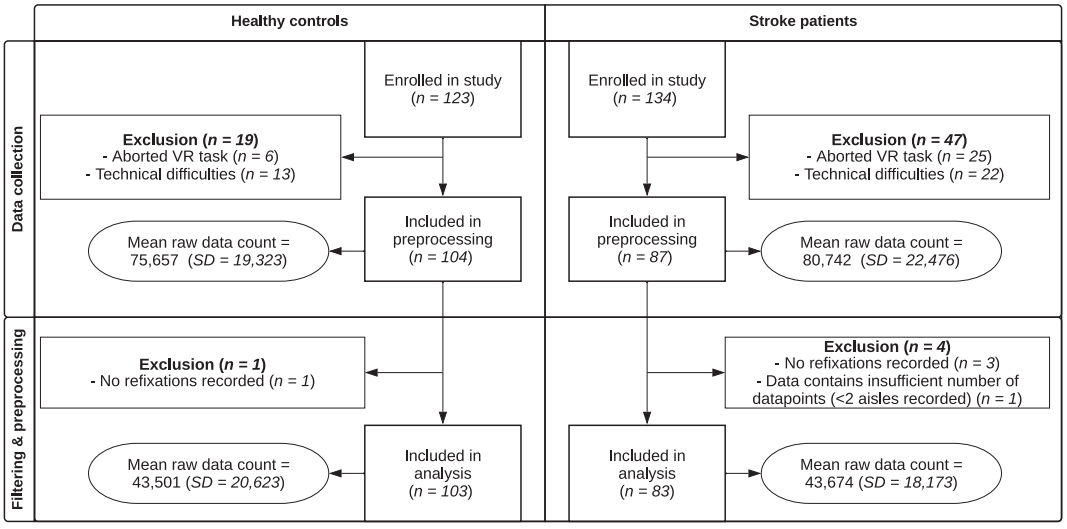

Figure 1. Flowchart of the data selection process with reasons for exclusion of participants specified. The mean data count indicates the mean number of datapoints per participant. Aborted virtual reality (VR) experiments include participants who experienced nausea, dizziness and/or fatigue, and were not able to finish the task. Technical difficulties include problems with data encoding leading to corrupted datafiles, which were excluded.

Table 2. Whole-task descriptives.

	Stroke patients	Healthy controls	Both groups
n			
Short list	62	27	89
Long list	21	76	97
Both tasks	83	103	
Task duration (minutes)			
Short list	15.45 (4.46)	11.42 (3.68)	14.23 (4.63)
Long list	13.70 (3.09)	14.24 (2.86)	14.13 (2.92)
Both tasks	15.01 (4.23)	13.50 (3.34)	
Effective sampling rate (after filtering)			
Short list	41.99 (17.04)	26.55 (4.58)	37.31 (16.10)
Long list	52.54 (15.33)	56.04 (15.78)	55.28 (15.75)
Both tasks	44.66 (17.25)	48.31 (18.90)	
Number of aisles visited			
Short list	10.19 (4.76)	7.93 (3.64)	9.50 (4.57)
Long list	7.64 (2.51)	9.13 (2.41)	8.80 (2.51)
Both tasks	9.53 (4.44)	8.81 (2.84)	
Number of products found			
Short list	2.18 (1.28)	2.39 (0.72)	2.25 (1.14)
Long list	3.48 (2.20)	4.43 (1.75)	4.21 (1.91)
Both tasks	2.52 (1.67)	3.89 (1.79)	

Table 3. Mean (*SD*) value of each feature over all participants, after aggregation by aisle and averaging over all aisles per participant. Values are reported per task type and per group.

	Stroke patients	Healthy controls	Both groups
n			
Short list	62	27	89
Long list	21	76	97
Both lists	83	103	
Fixations per second			
Short list	3.16 (0.40)	3.21 (0.38)	3.17 (0.39)
Long list	3.12 (0.19)	3.14 (0.24)	3.14 (0.23)
Both lists	3.15 (0.36)	3.16 (0.29)	
Median fixation duration (ms)			
Short list	172.16 (20.21)	177.08 (25.60)	173.66 (22.10)
Long list	181.68 (33.05)	164.62 (15.58)	168.41 (21.95)
Both lists	174.60 (24.51)	167.94 (19.57)	
Refixations per second			
Short list	0.15 (0.11)	0.14 (0.07)	0.15 (0.10)
Long list	0.14 (0.08)	0.15 (0.08)	0.15 (0.08)
Both lists	0.15 (0.10)	0.15 (0.08)	
Median time until refixations (s)			
Short list	3.38 (2.94)	2.46 (1.37)	3.10 (2.59)
Long list	1.85 (0.93)	2.80 (1.66)	2.59 (1.58)
Both lists	2.98 (2.66)	2.71 (1.60)	
Mean number of aisle revisits			
Short list	0.36 (0.34)	0.20 (0.26)	0.31 (0.33)
Long list	0.14 (0.14)	0.16 (0.15)	0.15 (0.15)
Both lists	0.31 (0.32)	0.17 (0.19)	

On average, 45.7% of data points were filtered out for the left eye and 46.3% were filtered out for the right eye. Figure 2 shows the distribution plots for the base features, split for task version (short-versus long list) and for participant group (stroke participants versus healthy controls).

3.2. Classification of task requirements: short versus long list

3.2.1. Including all participants

Over 50 independent runs, the Logistic Regression achieved an average AUC of .761 (*SD* = .06) and the Support Vector Machine achieved an average AUC of .753 (*SD* = .05; Figure 3, left-hand panel). Over 48 of those 50 independent runs, *mean number of aisle revisits* and *median time until refixations* are reported as having the highest median coefficients in non-negative terms (Figure 3, right-hand panel).

3.2.2. Including only healthy controls

Including only healthy controls (*n =103*), the Logistic Regression achieved an average AUC of .814 (*SD* = .07) and the Support Vector Machine achieved an average AUC of .801 (*SD* = .09). *Fixations per second* and *median time until refixations* are reported as having the highest median coefficients, reported over 49 of 50 independent runs.

3.2.3. Including only stroke patients

Including only stroke patients (*n = 83*), the Logistic Regression achieved an average AUC of .769 (*SD* = .08) and the Support Vector Machine achieved an average AUC of .743 (*SD* = .09). *Mean number of aisle revisits* and *refixations per second* are reported as having the highest median coefficients, reported over 45 of 50 independent runs.

3.3. Classification of groups: stroke patients versus healthy controls

3.3.1. Including both task versions

Over 50 independent runs, the Logistic Regression achieved an average AUC of .636 (*SD* = .07) and the Support Vector Machine achieved an average AUC of .642 (*SD* = .08; Figure 4, left-hand panel). Over those 50 independent runs, *mean number of aisle revisits* and *fixations per second* are reported as having the highest median coefficients in non-negative terms (although the coefficients of 20 runs are reported as they achieved an AUC >.65; Figure 4, right-hand panel).

3.3.2. Including only the short list

Including only data of the 3 product task (*n = 89*), the Logistic Regression achieved an average AUC of .634 (*SD* = .11) and the Support Vector Machine achieved an average AUC of .572 (*SD* = .15). *Fixations per second and mean number of aisle revisits* are reported as having the highest median coefficients, reported over 19 of 50 independent runs.

3.3.3. Including only the long list

Including only data of the 7 product task (*n = 97*), the Logistic Regression achieved an average AUC of .695 (*SD* = .11) and the Support Vector Machine achieved an average AUC of .623 (*SD* = .21). *Fixations per second* and *refixations per second* are reported as having the highest median coefficients, reported over 34 of 50 independent runs.

3.4. Correlation of base features

Testing for correlations between the base features revealed that the number of *fixations per second* and the number of *refixations per second* correlate, *r* = .18 (see Figure 5). This slight correlation is expected since refixations are included in the definition of all fixations. Secondly, the *median fixation duration* negatively correlates with *fixations per second* (*r* = -.37) and *refixations per second* (*r* = -.15). This is expected since the duration of fixations, combined with saccade- and blink durations, determine how many (re)fixations can be made within a second. Additionally, the *median time until refixations* correlates with *refixations per second* (*r* = .35). This surprising finding suggests that, when refixations are made more frequently, the re-fixated location is not likely to have been visited recently.

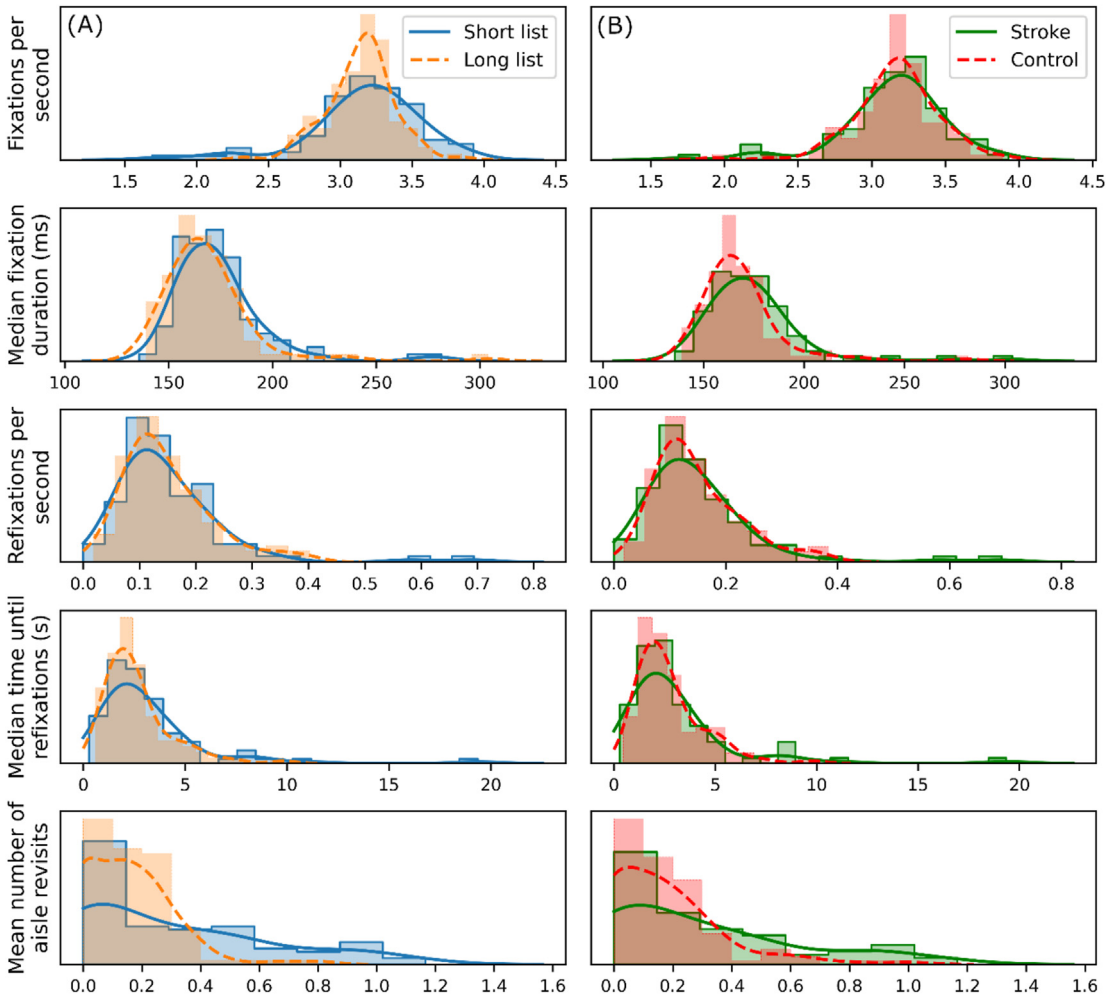

Figure 2. Distribution density plots of fixations per second, mean fixation duration (ms), refixations per second, mean time until refixations (s), and mean number of aisle revisits per visited aisle; split for task version (short-versus long list) in the left-hand panels [A]; and split per group (stroke versus control) in the right-hand panels [B]. Since these variables cannot contain negative values, kernel density estimations are cut-off at 0 and as such solely serve an illustrative purpose.

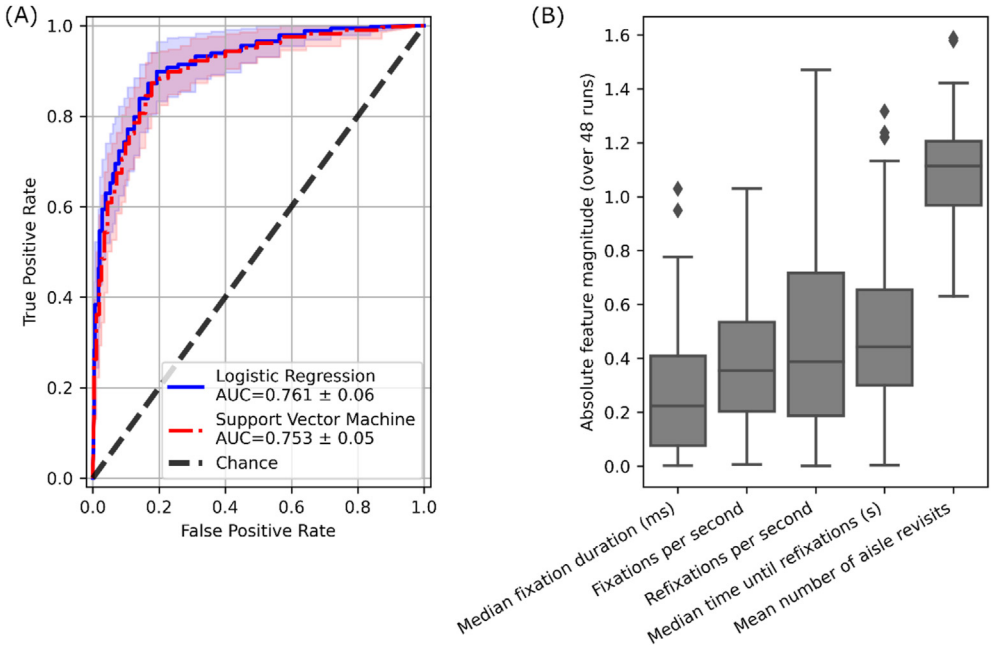

Figure 3. Results of classifying task requirements: short-versus long list (3 versus 7 products), with all participants included. Left-hand panel [A]: Averaged ROC curves for both models over 50 independent runs, with the shaded areas signifying ±1 SD from the mean. The dashed black line indicates where the ROC curve would lie if a model performed at chance level (AUC = .50). Right-hand panel [B]: Box-and-whisker plot of non-negative coefficients as extracted from 50 independent Logistic Regression model runs.

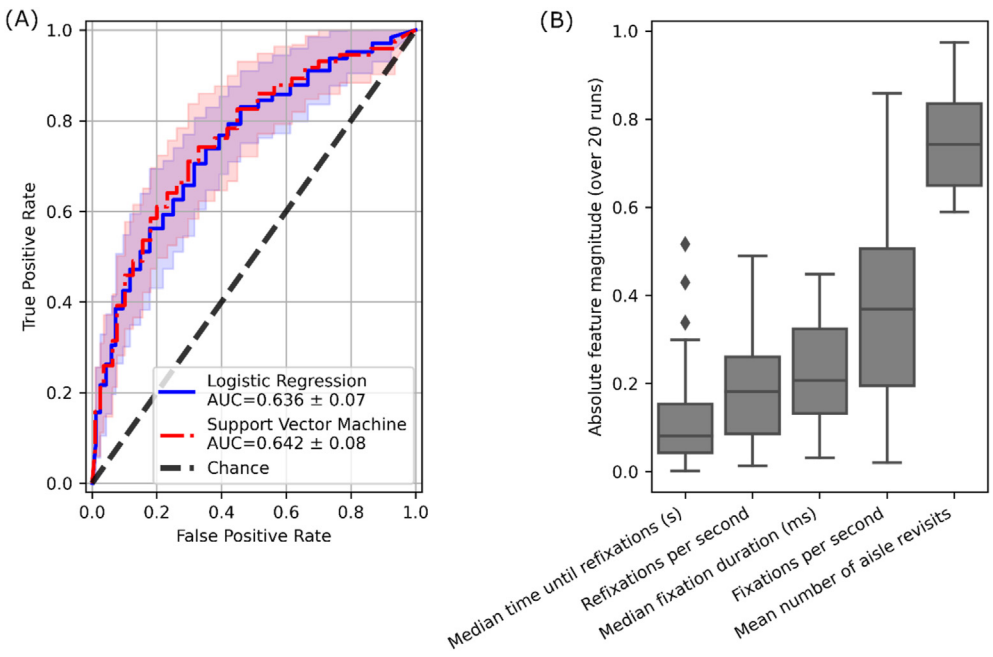

Figure 4. Results of classifying groups: stroke patients versus healthy controls, with all participants included. Left-hand panel [A]: Averaged ROC curves for both models over 50 independent runs, with the shaded areas signifying ±1 SD from the mean. The dashed black line indicated where the ROC curve would lie if a model performed at chance level (AUC = .50). Right-hand panel [B]: Box-and-whisker plot of non-negative coefficients as extracted from 42 of 50 independent Logistic Regression model runs.

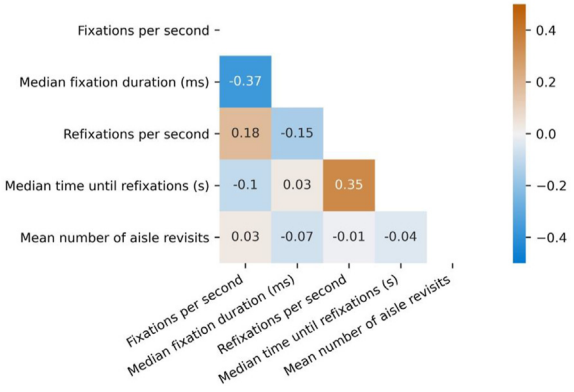

Figure 5. Correlation matrix for all base features, disregarding group or task type. The local base features were first aggregated by calculating the mean across aisles within each participant.

3.5. Logistic Regression based on eye tracking data as a cognitive screener

We investigate how well our Logistic Regression model, based on the VR eye tracking data, could perform as a screener for cognitive impairment as compared to the MoCA screener. Firstly, we identified whether the model correctly classified a patient in at least 50 percent of instances in which that patient was encountered in the test set across 50 model runs. We then identified whether that patient also suffered from symptoms of cognitive impairment as defined by a MoCA score <26 (Dautzenberg and de Jonghe, 2004; Nasreddine et al., 2005).

Of 76 patients for which MoCA scores were available, 22 patients performed below cut-off (MoCA score <26) and were correctly classified by the model at least 50% of the time. In the case of 21 patients, neither of those criteria were met. Therefore, we conclude that the VR eye tracking model and MoCA screener showed convergent outcomes on 43 patients (56.6% of the available sample). Conversely, 15 patients matched the eye tracking model's criterion but showed scores >26 on the MoCA, and 18 patients matched the MoCA criterion but not that of the eye tracking model. As such, we state that the VR eye tracking model and MoCA showed divergent outcomes on 43.4% of the available sample.

4. Discussion

The aim of the current study was to integrate eye tracking in a VR simulation and determine which eye movement features could best categorise between a short- and long shopping list, and between stroke patients and healthy control participants, using machine learning methods. Stroke patients and healthy control participants went shopping for either 3 or 7 products (i.e., short- and long list, respectively) in a VR supermarket environment while their gaze behaviour was continuously tracked. Using eye movement features, a model was able to correctly predict whether participants were assigned the short- or long list (task type) with an accuracy of .76 across stroke patients and healthy controls combined. Separated by group, task type could be classified with an accuracy of .81 in healthy controls, but slightly less so in stroke patients, with an accuracy of .77. Categorising whether the participant was a stroke patient or a healthy control was possible with an accuracy of approximately .64 with both task types combined. Separated by task type, stroke patients could be dissociated from healthy controls with an accuracy of .63 and .70 for the short- and long list, respectively.

We found that overall (1) the mean number of visited aisles that were revisited, and (2) the number of fixations per second were likely to be the most important features for distinguishing between the short and long lists, and for distinguishing between stroke patients and healthy controls. Since categorising tasks or groups based only on eye movement features is a complex issue, and machine learning algorithms generally rely on much larger sample sizes as discrimination becomes more complex, the current results show the clinical potential of the combined techniques of VR simulations, eye tracking, and machine learning.

Heterogeneity in demographic and stroke characteristics is usually quite large in large cohort studies, especially when including patients in different stages post-stroke. Moreover, in the present study, the severity and the *consequences* of stroke also varied considerably between patients, resulting in high variability in performance of the patient group during our virtual shopping task. As a consequence, performance of some patients with mild stroke or milder consequences of stroke will resemble performance of healthy controls. This could explain why the classification algorithm performed better at dissociating stroke patients from healthy control participants in the task in which 7 products had to be found as compared to in the 3 product task. Possibly, the short list does not dissociate well between healthy controls and stroke patients with

mild cognitive problems, as it might not assess the upper limits of cognitive functioning such that milder cognitive complaints will be undetected/unidentified. Although the accuracy of our model was relatively high, it is likely that performance of the model to categorise stroke patients and healthy control participants would have been higher if we would have applied more strict inclusion criteria with respect to stroke severity and consequences. However, given that our aim was to demonstrate the potential of the applied method in the general stroke population, we decided to apply relatively liberal inclusion criteria. Importantly, the small standard deviation of the AUC for both models in which all participants were included suggests that the performances of these models, based on our eye movement feature set, are relatively robust, and are not solely caused by outliers of some of the runs.

With respect to the generalisability, it is important to note that we included stroke patients admitted for either inpatient or (former) outpatient rehabilitation. In the Netherlands, stroke patients are referred for *inpatient* rehabilitation care when: (1) a safe discharge from hospital to home is not achievable within 5 days; (2) the patient is vital enough to participate in therapy; (3) a multidisciplinary approach is essential to reach complex rehabilitation goals; and (4) discharge from inpatient rehabilitation to home is expected in view of the prognosis and availability of the caregivers within 3 months. Stroke patients are referred for *outpatient* rehabilitation care when: (1) a safe discharge from hospital to home is achievable; and (2) a multidisciplinary approach is essential to reach rehabilitation goals. The average MoCA scores of stroke patients in our sample was 23.9, which suggests that they did have cognitive impairments, as the generally accepted cut-off point across clinical populations is 26/30 (Nasreddine et al., 2005). With the current inclusion criteria, the heterogeneity of the sample with respect to the severity of stroke and its consequences is therefore representative for the Dutch stroke population referred to a rehabilitation centre. For older or more severely hampered stroke patients (who are usually referred to geriatric rehabilitation centres), it is unclear how well the model can correctly identify them. We expect, however, that more strongly affected patients will exhibit more 'deviant' gaze behaviour, and thus it will be more likely that the model will correctly classify these patients. However, whether it is feasible to administer these type of tasks (e.g., VR simulations, complex search tasks in a dynamic environment) in patients with severe stroke currently remains unknown.

It is interesting to speculate on why both the *mean number of revisits per aisle* and the number of *fixations per second* were the most important features for categorising stroke patients and healthy controls. Stroke patients showed slightly fewer fixations per second than healthy controls, and there was a more pronounced tail in the distribution of mean number of aisle revisits compared to healthy controls (Figure 2). With respect to the *mean number of revisits per aisle*, the increased number of revisits in stroke patients could be due to the known inefficient search behaviour in stroke patients (e.g., Mark et al., 2004; Rabuffetti et al., 2012; Ten Brink et al., 2016b). In case a refixation is required (when not all required information at the re-fixated location is internalised during the previous aisle visit), the necessity of revisiting that aisle is increased, since information about that location might be lost. Indeed, it has been argued that the inefficient search strategy observed in stroke patients is due to memory deficits (e.g., Ten Brink et al., 2016a). Note that this would also explain the surprising positive correlation between median time till refixation and refixations per second. Specifically, when an aisle is revisited, multiple refixations per second can be made to items fixated on in the last visit to that aisle. Since the refixation due to a revisit to an aisle, the median duration until a revisit will be relative large. With respect to *fixations per second*, which is a variable which consists combinedly of the duration of fixations and the duration of saccades inbetween those fixations, it is known that the duration of a fixation is determined by the time required to process information at the fixated location, amongst other factors (Hooge and Erkelens, 1996; Salthouse and Ellis, 1980). Considering that difficulties in information processing are commonly reported in stroke patients (Gerritsen et al., 2003; Hochstenbach et al., 1998), and that stroke patients show inefficiencies in search behaviour which would lead to longer saccade durations, it is perhaps not surprising that the number of fixations per second is an important oculomotor feature to categorise stroke patients.

4.1. Limitations

Firstly, with respect to the patient sample, we could only report the most general characteristics related to stroke. As we included stroke patients from different centres, there was considerable variation in the clinical tests used for usual care. We only reported the stroke characteristics and the MoCA scores, as scores on other tests were only available for very small subsets. Future research is needed to further specify other potential subgroups based on clinical characteristics with systematically obtained stroke related characteristics (e.g., extensive neuropsychological assessment, severity of stroke in acute phase, lesion characteristics).

Secondly, as noted earlier, we included a relatively small sample. A larger sample size would provide more power for the model, allowing for building more complex cognitive models. We did, however, deliberately choose to use a Logistic Regression Model for its interpretability of feature importance. With more complex models, performance would likely be stronger, yet interpretation would be significantly more complex or not feasible.

A third limitation is related to our feature extraction. Although our custom extraction algorithm was designed to take data loss into account, the high variability in sampling rates (which we found to be as low as 25 Hz at times) and eye tracking quality (an average data loss of 46–47% after filtering) is likely to have led to imperfect feature extraction.

Lastly, we made use of data from different studies with different protocols (e.g., the VR assignment was either the only test or one of a variety of tests; the specific products on the shopping list differed between studies; etc.). Since only 22 of the stroke patients performed the task with 7 products, whereas 77 of the healthy control participants did, this might have been a confound in the classification of groups. Therefore, for each main analysis (i.e., the categorisation of the number of products and the categorisation of the groups), we ran secondary analyses in which we separately categorised either the number of products in each group, or the groups for each number of products. These categorising accuracies overlapped largely with those found in the main analysis, with the exception of dissociating between the short- and long task, where classification increased significantly with only healthy controls in the dataset and decreased significantly with only stroke patients in the dataset. Possibly, healthy controls adjust their behaviour more strongly than stroke patients when task difficulty changes (assuming that memorising and looking for 7 products is more difficult than doing so with 3 products). Although we conducted separate analyses for each category, the variation in total duration of the study - not just the VR assigment - might have had an influence on cognitive performance (e.g., fatigue) or motivation. Furthermore, specific products on the shopping list obviously impact the routes participants had to take, or the aisles with different target and distracting products. However, as we did not use the specific, detailed information of the eye movement features per aisle, but the collapsed data, we feel that the current results show that gaze behaviour in general can be used to categorise stroke patients and healthy control participants based on their overall eye tracking features.

5. Conclusion

The present model is a step towards a more sensitive assessment of impairments in visual search in a life-like situation. The positive results of the classification model are promising for application in patient diagnosis and rehabilitation. In the future, the model could assist in the early detection of abnormalities in cognition in a variety of diseases. Clearly, eye movement data contains a rich set of signatures for detecting cognitive deficits, opening the door to potential clinical applications.

Declarations

Author contribution statement

Veerle H.E.W. Brouwer, Isabel K. Gosselt: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

Sjoerd Stuit, Alex Hoogerbrugge, Antonia F. Ten Brink: Analyzed and interpreted the data; Wrote the paper.

Stefan Van der Stigchel, Tanja C.W. Nijboer: Conceived and designed the experiments; Wrote the paper.

Funding statement

This work was supported by HandicapNL under Grant [R2015010 and R201705758] to Tanja C.W. Nijboer, seed money grants by Focus Areas DataScience and Research IT from Utrecht University to Tanja C.W. Nijboer, and ERC [ERC-CoG-863732] to Stefan Van der Stigchel.

Data availability statement

The data that has been used is confidential.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

We would like to thank the UMC Utrecht and De Hoogstraat Rehabilitation for sharing the anonimised data. We thank Atoms2Bits for providing the VR software. A special thanks to all participants for their contribution. We would like to thank Timo Kootstra for his contributions to the model and data preprocessing code. We thank the rehabilitation physicians, (neuro) psychologists and occupational therapists for their help evaluating the stroke patients. We thank our students for their help collecting the data.

References

Borji, A., Itti, L., 2014. Defending Yarbus: eye movements reveal observers' task. J. Vis. 14 (3), 29.

Borji, A., Lennartz, A., Pomplun, M., 2015. What do eyes reveal about the mind? Neurocomputing 149 (PB), 788–799.

Bosch, S.E., Neggers, S.F.W., Van der Stigchel, S., 2013. The role of the frontal eye fields in oculomotor competition: image-guided TMS enhances contralateral target selection. Cerebr. Cortex 23 (4), 824–832.

Brooks, B.M., Rose, F.D., 2003. The use of virtual reality in memory rehabilitation: current findings and future directions. NeuroRehabilitation 18 (2), 147–157.

Carassa, A., Morganti, F., Tirassa, M., 2005. A Situated Cognition Perspective on Presence Affordances as Representations. eScholarsh.

Carette, R., Elbattah, M., Cilia, F., Dequen, G., Guérin, J.-L., Bosche, J., 2019. Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies, pp. 103–112.

Charron, C., Hoc, J.-M., Milleville-Pennel, I., 2010. Cognitive control by brain-injured car drivers: an exploratory study. Ergonomics 53 (12), 1434–1445.

Clay, V., König, P., König, S., 2019. Eye tracking in virtual reality. J. Eye Move. Res. 12 (1), 1–18.

Corbetta, M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Ollinger, J.M., Drury, H.A., Linenweber, M.R., Petersen, S.E., Raichle, M.E., Van Essen, D.C., Shulman, G.L., 1998. A common network of functional areas for attention and eye movements. Neuron 21 (4), 761–773.

Cyr, A.-A., Stinchcombe, A., Gagnon, S., Marshall, S., Hing, M.M.-S., Finestone, H., 2009. Driving difficulties of brain-injured drivers in reaction to high-crash-risk simulated road events: a question of impaired divided attention? J. Clin. Exp. Neuropsychol. 31 (4), 472–482.

Dautzenberg, P.L.J., de Jonghe, J.F.M., 2004. Montreal Cognitive Assessment: afname- en scoringinstructies.

Delazer, M., Sojer, M., Ellmerer, P., Boehme, C., Benke, T., 2018. Eye-tracking provides a sensitive measure of exploration deficits after acute right MCA stroke. Front. Neurol. 9 (JUN), 1–9.

Everling, S., Fischer, B., 1998. The antisaccade: a review of basic research and clinical studies. Neuropsychologia 36 (9), 885–899.

Gerritsen, M.J.J., Berg, I.J., Deelman, B.G., Visser-Keizer, A.C., Jong, B.M., 2003. Speed of information processing after unilateral stroke. J. Clin. Exp. Neuropsychol. 25 (1), 1–13.

Hochstenbach, J., Mulder, T., van Limbeek, J., Donders, R., Schoonderwaldt, H., 1998. Cognitive decline following stroke: a comprehensive study of cognitive decline following stroke. J. Clin. Exp. Neuropsychol. 20 (4), 503–517.

Hooge, I.T.C., Erkelens, C.J., 1996. Control of fixation duration in a simple search task. Percept. Psychophys. 58 (7), 969–976.

Husain, M., Mannan, S., Hodgson, T., Wojciulik, E., Driver, J., Kennard, C., 2001. Impaired spatial working memory across saccades contributes to abnormal search in parietal neglect. Brain 124 (5), 941–952.

Kassner, M., Patera, W., 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, pp. 1151–1160.

Kootstra, T., Teuwen, J., Goudsmit, J., Nijboer, T., Dodd, M., Van der Stigchel, S., 2020. Machine learning-based classification of viewing behavior using a wide range of statistical oculomotor features. J. Vis. 20 (9), 1.

Lagun, D., Manzanares, C., Zola, S.M., Buffalo, E.A., Agichtein, E., 2011. Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. J. Neurosci. Methods 201 (1), 196–203.

Mark, V., Woods, A., Ball, K., Roth, D., Mennemeier, M., 2004. Disorganized search on cancellation is not a consequence of neglect. Neurology 63 (1), 78–84.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., Dodd, M.D., 2011. Examining the influence of task set on eye movements and fixations. J. Vis. 11 (8), 17.

Najemnik, J., Geisler, W.S., 2005. Optimal eye movement strategies in visual search. Nature 434 (7031), 387–391.

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H., 2005. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53 (4), 695–699.

Parsons, T.D., 2015. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. Front. Hum. Neurosci. 9 (DEC), 1–19.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (January), 2825–2830.

Pupil Labs, 2020. Pupil Labs HTC Vive Binocular Add-On. October.https:pupil-labs.com/.

Pusiol, G., Esteva, A., Hall, S.S., Frank, M., Milstein, A., Fei-Fei, L., 2016. Vision-based classification of developmental disorders using eye-movements. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer, pp. 317–325.

Rabuffetti, M., Farina, E., Alberoni, M., Pellegatta, D., Appollonio, I., Affanni, P., Forni, M., Ferrarin, M., 2012. Spatio-temporal features of visual exploration in unilaterally brain-damaged subjects with or without neglect: results from a touchscreen test. PLoS One 7 (2), e31511.

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

Rizzo, A.A., Schultheis, M., Kerns, K.A., Mateer, C., 2004. Analysis of assets for virtual reality applications in neuropsychology. Neuropsychol. Rehabil. 14 (1–2), 207–239.

Rizzolatti, G., Riggio, L., Dascola, I., Umiltá, C., 1987. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. Neuropsychologia 25 (1), 31–40.

Salthouse, T.A., Ellis, C.L., 1980. Determinants of eye-fixation duration. Am. J. Psychol. 93 (2), 207.

Ten Brink, A.F., Biesbroek, M.J., Kuijf, H.J., Van der Stigchel, S., Oort, Q., Visser-Meily, J.M.A., Nijboer, T.C.W., 2016a. The right hemisphere is dominant in organization of visual search—a study in stroke patients. Behav. Brain Res. 304, 71–79.

Ten Brink, A.F., Van der Stigchel, S., Visser-Meily, J.M.A., Nijboer, T.C.W., 2016b. You never know where you are going until you know where you have been: disorganized search after stroke. J. Neuropsychol. 10 (2), 256–275.

Van der Stigchel, S., Hollingworth, A., 2018. Visuospatial working memory as a fundamental component of the eye movement system. Curr. Dir. Psychol. Sci. 27 (2), 136–143.

Verhage, F., 1965. Intelligence and age in a Dutch sample. Hum. Dev. 8 (4), 238–245.

Walle, K.M., Nordvik, J.E., Becker, F., Espeseth, T., Sneve, M.H., Laeng, B., 2019. Unilateral neglect post stroke: eye movement frequencies indicate directional hypokinesia while fixation distributions suggest compensational mechanism. Brain Behav. 9 (1), 1–21.

You, S.H., Jang, S.H., Kim, Y.-H., Hallett, M., Ahn, S.H., Kwon, Y.-H., Kim, J.H., Lee, M.Y., 2005. Virtual reality–induced cortical reorganization and associated locomotor recovery in chronic stroke. Stroke 36 (6), 1166–1171.