Eye Tracking Data Seahaven

a summary of the processing stages by Jasmin Walter

the most recent version can be found here: https://github.com/jasminwalter/NBP-VR-Eyetracking

RUNNING THE EXPERIMENT

→ different types of data are created during the runtime of the experiment and stored in the following folders:

- EyeBoxPos
  - ➞ contains EyeBoxPox .txt files for every participant
- EyesOnScreen
  - ➞ contains EyesOnScreen .txt files for every participant
- Position
  - ➞ contains Position .txt files for every participant
- PupilRecording
  - ➞ contains a folder for every participant
  - ➞ Folder contains raw data of pupil labs eye tracker (200 hz sampling rate)
  - ➞ Can be opened and exported to csv files with Pupil Player (plugin “raw data exporter”)
- Validation
  - ➞ contains several Validation .txt files for every participant

→ According to Viviane Kakerbeck Script Overview, the data types contain: (https://github.com/vkakerbeck/NBP-VR-Lab)

File Name	Data	Script
EyesOnScreen_VP#	2D coordinates of gaze (normalized) = **(CenterX,CenterY)** or **(0.000000, 0.000000)**	PupilGazeTracker
EyeBoxPos_VP#	3D coordinates of box position: **(x,y,z)**	PupilGazeTracker
Positions_VP#	**(x,y,z,rx,ry,rz,timestamp (in sec),PupilTimeStamp)**	Recorder
Validation2D_#_NumVal	**Degree of error for each point, avg, time, last cal, error in x and y dir (+avg)**	PupilGazeTracker

→ Therefore, as I understand it:

- EyesOnScreen – contains 2D coordinates of eye gaze position (about 30 Hz sampling rate)
- EyeBoxPos – contains 3D coordinates of vector in gaze direction calculated during runtime of the experiment (about 30 Hz sampling rate)
- Position – contains coordinates (x,y,z and rotation information) of eye tracking headset. (about 30 Hz sampling rate)
  - ➞ Plus it contains the timestamp calculated during runtime, which is used for the synchronization and ray casting in the DrawViewingPath script
  - ➞ It also contains the pupil labs timestamp – however it cuts of most of the decimal digits of the timestamp, therefore the timestamp cannot be differentiated anymore
    (Consequently, the pupil labs timestamp cannot be used to synchronize the VR headset information with the raw Pupil Labs eye tracking data)

NOTE:

→ The position files contain 2 kinds of errors

- Sometimes a line is doubled (or tripled)
- Sometimes the timestamp based on seconds (7th column) is incorrect. The rest of the line appears to be correct (no doubled values), while the sec timestamp is the same as the one in the line before.

→ I am not sure, how these errors affect the ray casting and synchronization, consequently, how it affects the ViewedHouses and Heatmap3D data.

Ray casting process =
running script ViewDrawingPath

→ After experiment the script ViewDrawingPath or DrawAllPaths is run in Unity. It synchronizes the “eyeboxpos” files with the “position” files and calculates the ray cast to determine the hit points. Running the script produces two types of data saved in the following folders:

- ViewedHouses
  - ➞ Contains ViewedHouses .txt files for every participant
- Heatmat3D
  - ➞ Contains 3DHeatmap .txt files for every participant

NOTE:

!!! There used to be an error in the ViewDrawingPath and DrawAllPath script that messed up the synchronization and lead to shorter ViewedHouses and Heatmap3D files – in other words, the raycasting process did not cover the duration of the full experiment and hence, created files for only part of the experiment. The error was fixed in February 2019 by Nicolas Kuske (nkuske@uos.de), therefore

!!!! no ViewedHouses file created BEFORE FEBRUARY 2019 should be used in further analysis!!!!

→ According to Viviane Kakerbeck Script Overview, the files created when running the ray casting script contain:

ViewedHouses_VP#	**HouseViewed, distance, timestamp (sec. since start)**. If HouseViewed=NH: if distance=0: eye detected with confidence <0.5 if distance=200: no object hit (eg. Sky) else: other object than house hit	DrawViewingPath
Heatmap3D(RandomGaze/ RandomPos)_VP#	List of **(x,y,distance)** of fixations during the session	DrawViewingPath

→ In other words:

→ The viewedHouses file contains data points in a 33,33334 ms interval. Each data point contains information about the object the ray cast has hit (house, sky or noHouse category that includes everything in Seahaven that is not a house or the sky (e.g. road, grass, trees, sea...)), the distance between hit point and VR headset, as well as the timestamp that was created during runtime of the experiment.

→ Note the following code

- NH object in combination with distance 0 – denotes that the data point carries no valid information = the pupil was detected with a probability of less than 50% and no information about the ray cast is available
- NH object in combination with distance 200 – means that the hit point of the ray cast was the sky (= no object was hit)

→ The data points in viewedHouses files are comparable to a 30 Hz sampling rate

→ Sometimes a high percentage of data points do not carry valid information, therefore, be careful whether to keep participants in the analysis!

Up to here, I only summarized my knowledge about the process and processing levels of the data created in the Seahaven project, before I started working with the eye tracking data. The following section and information is based on my work and my scripts.
All my scripts and related information can be found here:
https://github.com/jasminwalter/NBP-VR-Eyetracking

Cleaning and Preprocessing of ViewedHouses files

by Jasmin Walter

→ The following cleaning and preprocessing information only refers to the ViewedHouses files.

Note: there are currently 2 versions of pre-processing pipelines:

- Version 2 – used for the paper on graph theory
- Version 1 – used in my bachelor thesis (and the work of the people who used any of my scripts written before October 2019)

Cleaning and Preprocessing Version 2

Note: if you use the cleaning and preprocessing pipeline for the viewedHouses files, make sure all scripts have the identifier '*scriptname*_**V2**.m'

Script condenseViewedHouses_V2.m

→ Cleans and pre-processes the ViewedHouses files based on the specified participant list, in detail the script:

- Renames NH in combination with distance 200 to "sky"

- Renames all data samples, that do not carry valid information (NH + distance 0) to “noData”
- Creates overview of complete amount of data rows and percentage of “noData” rows
- Reorganizes the file structure, such that consecutive data rows with hit points onto the same object, are combined into one row, and the consecutive amount of samples are listed.

File name after running script: _condensedViewedHouses.mat

ViewedHouses.txt

32	014_0	12.64634	1.02
33	014_0	14.00599	1.06
34	014_0	14.04966	1.09
35	NH	26.04359	1.12
36	NH	26.04358	1.15
37	NH	26.55452	1.19
38	NH	28.08442	1.22
39	066_0	28.69236	1.25
40	066_0	28.69236	1.29
41	066_0	28.90836	1.32
42	066_0	29.16993	1.35
43	NH	0	1.39
44	NH	0	1.42
45	NH	0	1.45
46	002_0	31.31488	1.48
47	NH	200	1.52
48	NH	200	1.55
49	NH	200	1.58

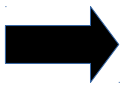


Table created after running condenseViewedHouses_V2.m

	1 House	2 Time	3 Samples	4 Distances
	'014_0'	533.3333	16	12.8154
	'NH'	133.3333	4	26.6815
	'066_0'	133.3333	4	28.8658
	'noData'	100	3	0
	'002_0'	33.3333	1	31.3149
	'sky'	233.3333	7	200
	'noData'	33.3333	1	0
	'sky'	133.3333	4	200
	'015_0'	233.3333	7	49.5763
	'noData'	100	3	0
	'050_0'	233.3333	7	18.9926
	'148_6'	133.3333	4	14.7177
	'noData'	466.6667	14	0

→ Meaning of the respective columns in the new CondenseViewedHouses files

- House – denotes the object hit during ray casting, can contain a house number, sky, NH category or noData samples
- Time – combined time of all consecutive data points hitting that object (time unit = milliseconds)
- Samples – amount of consecutive data points combined into one row
- Distance – mean viewing distance of all consecutive data points combined into one row (distance unit = Unity® units that are comparable to meters)

Script
cleanParticipants_V2.m

→ Uses overview created in condenseViewedHouses_V2.m, specifically the ratio of "noData" rows compared to the total amount of data rows. Creates a new participant list, that has only participants listed who had less than the specified percentage of "noData rows" (noData rows designate rows with no pupil information, hence no viewing information). The threshold is currently set to 30% of removed data during cleaning, consequently, only participants with less than 30% removed data are listed in the new participant list.

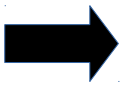
Script
interpolateLostData_V2.m

→ Interpolates "noData" samples iff they qualify with the following criteria
- Amount of consecutive noData samples to interpolate is 0-7 samples
- The noData samples are between the same house

→ In this case, the no data samples are interpolated with the same house name as the house before and after

If the houses before and after are the **same** house - **interpolate**

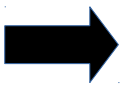
1 House	2 Time	3 Samples	4 Distances
'008_0'	133.3333	4	15.8389
'noData'	33.3333	1	0
'008_0'	33.3333	1	16.9240



1 House	2 Time	3 Samples	4 Distances
'008_0'	200	6	15.8389

if the houses before and after are **different** houses – **no interpolation**

1 House	2 Time	3 Samples	4 Distances
'008_0'	66.6667	2	27.4578
'noData'	100	3	0
'108_3'	33.3333	1	40.5254



NO INTERPOLATION!!!

File name after running script: _interpolatedViewedHouses.mat

Script
gazes_vs_noise_V2.m

- → Divides the data into two files based on the number consecutive of data points hitting the same object. If an object has more than 7 consecutive data points, it is assumed be be gazed at by the participant and used in further analysis. Consequently, only objects that were consecutively looked at for more than 233,333 ms are used in further analysis.
- → Creates 2 type of data files
  - gazes_data.mat files
    - ➞ Data points with more than 7 consecutive samples – used for further analysis
  - noisy_data.mat files
    - ➞ Data points with 7 or less than 7 consecutive samples – excluded from further analysis

File names after running script: gazes_data.mat & noisy_data.mat

Script
CreateGraphs_V2.m

- → Creates graph objects from gazes_data .mat files (that were created by script: gazes_vs_noise_V2.m)
- → Only uses data information about houses (sky and NH category are excluded)
- → A node in the graph corresponds to a house in Seahaven
- → If an edge exists between two nodes, the participant made one direct visual connection from one house to another during the experiment. In other words, the participant constitutively looked from one house to another, that is represented by a connection (edge) between the respective nodes in the graph
- → Information about how often a specific visual connection between houses occurred during the experiment is excluded. The graph only shows that the connection was made at least one time
- → All connections to noData samples are excluded from the graph, therefore some nodes might not be connected to the rest of the graph

NOTE: all remaining noData samples do not lead to connections in the graph, hence all noData samples disconnect the sample before and after the noData sample in the graph.

File name after running script: _graph.mat

Cleaning and Preprocessing Version 1

Version1 = Cleaning and preprocessing of the eye tracking data used in my bachelor thesis and by others who build their work on my scripts I wrote in my bachelor thesis. All scripts can be found here: https://github.com/jasminwalter/NBP-VR-Eyetracking/tree/master/Bachelor%20Thesis

Note: if you use the cleaning and preprocessing pipeline of the version 1 for the viewedHouses files, make sure all scripts have the identifier '*scriptname*_**V1**.m'

Cleaning and Preprocessing of ViewedHouses files

by Jasmin Walter

→ My Bachelor thesis and analysis of the eye tracking data was based on the ViewedHouses Files. Consequently, the following cleaning and preprocessing only refers to the ViewedHouses files.

Script
condenseViewedHouses_V1.m

→ Cleans and pre-processes the ViewedHouses files based on the specified participant list, in detail the script:

- Renames NH in combination with distance 200 to "sky"
- Removes all data samples, that do not carry valid information (NH + distance 0)
- Creates overview of complete amount and percentage of removed data samples

- Reorganizes the file structure, such that consecutive data rows with hit points onto the same object are combined into one row and consecutive amount of samples are listed.

ViewedHouses.txt

069_0	67.9864	0.59
069_0	69.67004	0.63
069_0	65.83246	0.66
069_0	69.38956	0.69
069_0	74.02974	0.73
NH	127.6808	0.76
069_0	75.27911	0.79
NH	126.2815	0.82
NH	112.2871	0.86
NH	19.18605	0.89
NH	19.58365	0.92

Table created after running condenseViewedHouses.m

1 House	2 Time	3 Looks	4 Distances
'069_0'	166.6667	5	69.3816
'NH'	33.3333	1	127.6808
'069_0'	33.3333	1	75.2791
'NH'	333.3333	10	41.6504
'069_0'	100	3	64.6198

-

→ Meaning of the respective columns in the new CondenseViewedHouses files

- House – denotes the object hit during ray casting, can contain a house number, sky, or NH category
- Time – combined time of all consecutive data points hitting that object (time unit = milliseconds)
- Looks – amount of consecutive data points combined into one row
- Distance – mean viewing distance of all consecutive data points combined into one row (distance unit = Unity® units that are comparable to meters)

Script
cleanParticipants_V1.m

→ Creates a new participant list, that has only participants listed who had less than the specified percentage of their eye tracking data removed during cleaning (script “condenseViewedHouses.m”). The threshold is currently set to 30% of removed data during cleaning, consequently, only participants with less than 30% removed data are listed in the new participant list.

Script
fixated_vs_noise_V1.m

→ Divides the data into two files based on the number consecutive of data points hitting the same object. If an object has more than 7 consecutive data points, it is assumed be be fixated by the participant and used in further analysis. Consequently, only objects that were consecutively looked at for more than 233,333 ms are used in further analysis.

→ Creates 2 type of data files

- Fixated_objects .mat files
  - ➞ Data points with more than 7 consecutive samples – used for further analysis
- noisy_data .mat files
  - ➞ Data points with 7 or less than 7 consecutive samples – excluded from further analysis

Script
CreateGraphs_V1.m

→ Creates graph objects from fixated_objects .mat files (that were created by script: fixated_vs_noise.m)

→ Only uses data information about houses (sky and NH category are excluded)

→ A node in the graph corresponds to a house in Seahaven

→ If an edge exists between two nodes, the participant made one direct visual connection from one house to another during the experiment. In other words, the participant constitutively looked from one house to another, that is represented by a connection (edge) between the respective nodes in the graph

→ Information about how often a specific visual connection between houses occurred during the experiment is excluded. The graph only shows that the connection was made at least one time