Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Insights from a very large EEG+Eyetracking dataset

Okan Mazlum

Studiengang: Informatik

Prüfer/in: Jun. -Prof. Dr. Benedikt Ehinger

Betreuer/in: Dr. Jevri Hanna

Beginn am: October 13, 2025

Beendet am: April 13, 2026

Abstract

sample abstract [1]

Introduction.

---

[1] sample footnote

Contents

List of Figures

List of Tables

Listings

1 Introduction

... [proper intro]

1.1 Datasets

...... [proper explaination of what the purposes and differences of the HBN and HBN-EEG datasets are]

1.1.1 Paradigms

- ...

- **Symbol Search:** [quick explanation of symbol search paradigm. it can be seen how the subject moves their gaze horizontally line by line. the image captures only the first page, after pressing a button, this gaze pattern repeats (starting at the top left). i wanted to include this image because i think it nicely visualizes the et data that will be analyzed (very horizontal-heavy eye movement, shows types of saccades that are performed, new stimulus onset + large saccade with every new page)]
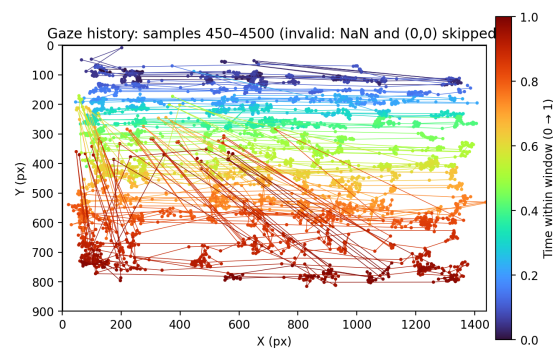

Figure 1.1: Gaze history visualization [Clean up this image and maybe find a better subject]

2 Related Work

. . .

3 Data and Methods

3.1 Data acquisition and sources

[The part on datasets in the Introduction covers the datasets on a basic level, here some more in-depth technical description is needed to aid in understanding the next sections. this also needs more detail]

HBN-EEG dataset

The dataset was retrieved from NeMAR [1] and consists of 11 `.zip` archives totalling 1.7 TiB, each containing curated EEG data in BIDS format.

HBN dataset

The original HBN dataset is hosted on the public fcp-indi S3 bucket [2]. The folder `data/Archives/HBN/EEG/` contains 4576 `.tar.gz` subject archives totalling 5.6 TiB. Each subject archive holds at most 3 Folders, `Behavioral`, `EEG` and `Eyetracking` [double check logs whether subjects exist where not all 3 are present].

- **Behavioral:** Contains phenotypic data about the subject in at most two different formats, `.csv` and `.mat`.

- **EEG:** Contains preprocessed [2] and raw EEG data. Since the HBN-EEG dataset will be used over the original HBN dataset for EEG data, this folder is of no further interest.

- **Eyetracking:** Contains at most 4 subfolders: `idf`, `txt`, `mat` and `tsv` [actually check whether subjects exist where all 4 are present]. Each folder represents the same eyetracking data in a different format. [go into a little more detail here, specifically that tsv ET does not contain necessary events to process it, that the idf folder is always empty, and that mat requires conversion to .txt]

---

[2]The exact preprocessing steps that were applied are unknown and the preprocessed data is not meant to be used `https://www.nitrc.org/forum/forum.php?forum_id=10003&thread_id=15454`

Based on the release documentation [3], updates between releases mostly concern data availability and dataset curation (e.g., additions of participants, corrections to phenotypic tables or metadata). In particular, there are no changes that would require release-specific treatment of the EEG data. For this reason, EEG data from the full HBN-EEG collection were merged together to form a single integrated BIDS dataset. This enables a single, consistent preprocessing and analyis pipeline across all subjects. This merged dataset was then enriched with the associated eye-tracking and phenotypic data for each subject from the original HBN archives. The integration of this type of data is not yet standardized in BIDS, and this step will cause BIDS validators [4] to consider the dataset as invalid. The tools used in this study (mne-bids-pipeline and Unfold.jl) remained compatible with this developmental format, simply ignoring the extra files that are not of relevance. The following sections cover how this merged dataset was created.

3.1.1 Download

Custom helper scripts were used to download and unpack all required source archives in a restartable manner. This was done to improve the reproducibility of the study compared to manual downloading. Downloads were performed with parallel workers and to ensure integrity of the files, incomplete transfers were written to a staging directory and only moved to the final location after the expected file size was reached. This approach was taken both for the retrieval from the fcp-indi S3 bucket and for the web download from NeMAR. After download, all archives were extracted using dedicated unpacking scripts making use of the same paradigms to avoid partial extractions and make use of parallelization.

3.1.2 Merging

Finally, further scripts make use of the downloaded files to deterministically create the final merged BIDS dataset. For files that are required as-is in the new merged dataset, e.g. EEG and eye-tracking recordings, the script creates symlinks to the downloaded files and renames them to be BIDS conform. Other files like the `participants.tsv` or other metadata files for the merged dataset have to be generated by merging metadata files from the separate HBN-EEG releases (i.e. concatenating all `participants.tsv` or some summary tables together into one file). To support traceability, some files were also augmented with their source release numbers. For files that were contained in all releases with only minor differences irrelevant to the merged result, the file from the first release was chosen.

Because metadata conventions and file naming practices vary slightly across releases, some harmonization steps were applied during dataset construction. These include:

1. Standardization of BIDS file names (subject identifiers, task labels, run indices, and separators)

2. Completion and canonicalization of `*_channels.tsv` contents by adding required columns and enforcing consistent channel types and units

3. Normalization of tabular files by harmonizing delimiters, column names, and column order

4. Targeted handling of a small number of subjects (< 20) with irregularities beyond these standard cases (e.g. by hardcoded renaming, conversion, or exclusion).

Throughout the merging process, the scripts validated expected directory structure and file contents, logging and/or terminating on any unexpected finds.

3.1.3 Synchronization with eye-tracking data

[reformulate and majorly elaborate on these paragraphs, depending on what will be written in the introduction] To synchronize EEG data with eye-tracking data, a fork of the mne-bids-pipeline was used, which is currently being developed for this exact purpose. It works by lining up shared events in the EEG and eye-tracking recording. The outputted `.fif` files contain eye-tracking samples and channels alongside the usual EEG data.

Due to the number of shared events being the only grounds for synchronization accuracy, the movie watching paradigms in the dataset were unsuitable for this study, containing only 2 shared events per recording. This only left the symbol search paradigm 1.1.1 as a candidate to analyze [depending on how quality scoring turns out, some movie recordings will likely still be used if deemed to have good synchronization despite small shared event count].

[@jevri if i have a movie watching run, and all quality metrics look good, is it fine to still consider it for analysis even though it only has two shared events? or should i blanket ban all recordings with shared events under some treshold? i feel like if the xcorr curve looks good, it still might be worth considering]

3.1.4 Parsing of eye-tracking data

Eye-tracking data in the HBN dataset were recorded using the iView-X Red-m by SensoMotoric Instruments [5]. These recordings usually come in a proprietary `.idf` format, but were converted to `.txt` and `.tsv` for almost all subjects [give exact number. maybe mention that despite the converter being obsolete/abandonware, there seem to exist alternatives for conversion. however no .idf files remain in the dataset so this is irrelevant].

An eye-tracking recording has two files associated with it.

19

- Samples file (`sub-<id>_task-symbolSearch_et.txt`): [briefly describe structure of this file] [essentially, the file starts with some headers containing a few useful pieces of information, followed by a tsv (with slight quirks). this table lists all ET samples as well as all user events with their respective timestamps]

- Events file (`sub-<id>_task-symbolSearch_et_Events.txt`): [briefly describe structure of this file] [again, it's some headers followed by a quasi-tsv. this time the tsv contains all user events (these were confirmed to be fully identical to the user events in the samples file) and saccade, blink, and fixation events. these events include fixation duration, position, etc...]

Subjects may only have an Events file without Samples file [how many subjects have this property?] in which case an analysis is still possible to some degree, as detailed gaze positions are not required for all analyses performed in this study. Similarly, for subjects that have a Samples file without an Events file [how many subjects have this property?], saccade and fixation information can be recreated from eye-tracking information [this is not currently done]. Only subjects with both files missing [how many subjects?] are skipped.

The used mne-bids-pipeline fork exclusively supported eye-tracking files in eyelink format (`.asc`, `.edf`), and had to be extended with the functionality to parse converted `.idt` files. For later curation of select subjects, numerous metrics were additionally collected during the parsing and synchronization process [this step has to be done because manual curation is not feasible in a dataset of this size]. As this step is ultimately very specific to this study, a second variant of the modified mne-bids-pipeline was created that exclusively focused on adding support for `.idt` files.

[quick summary of `read_raw_iview`. formulate this properly]
[[if samples file is present]]
[read sample rate from header]
[locate start of tsv, special treatment for time and type cols]
[remove channels that only have nan values]
[count broken samples, i.e. sample has almost all gaze data == 0 (maybe omit as it's mentioned later in metrics)]
[parse user events (type == ̈MSG ̈) into mne Annotations]
[estimate actual sampling frequency from times channel, as otherwise some annotations may fall outside of the raw time range]
[add everything to Raw object]

[if events file is present]
[parse user events (type == ̈UserEvent ̈) into mne Annotations]
[parse Fixation, Blink, Saccade events into mne Annotations. Make use of ̈extras ̈attributes that can be added to Annotations since MNE-Python v1.10.0 to add the auxiliary info like fixation position, saccade speed, ...]

. . . . . .

[misc hurdles to maybe mention]

[mne internal function `_combine_annotations` (concatenating a tuple of annotations) seems to not support ëxtras. when attempting, the resulting combined annotation object is missing all ëxtrasïnfo. to fix this, the internal function was adjusted]

[some ET `.txt` files contained special symbols that crashed mne when trying to name the channels, so they had to be renamed]

[adjusted prefixing events with `ET_`, which previously caused event messages to be tructuated. look at code for this again, it might have slightly modified functionality]

[when texttttfirst_samp differs between streams, shift back EEG annotations so they are synced. without this step, some annotations may again fall out of range] [added assert that at least 2 sync events are needed, as regression crashes otherwise]

3.1.5 Capturing of synchronization quality metrics

Throughout the parsing and synchronization process, metrics are collected to assess the quality of the input EEG and Eye-tracking files, as well as the resulting synchronized file. These metrics include meta properties (sample counts, sampling rates, channel counts), basic checks (number of invalid samples, average saccade, fixation and blink properties) and previously built-in synchronization quality feedback (regression slope/intercept, residual timing error with counts within ±1 and ±4 EEG samples, cross-correlation between HEOG and screen gaze position). For a full list of metrics, see the table in the Appendix [todo].

Once finished, the script writes all metrics as a as `*_metrics.json` into the derivatives folder next to the standard mne-bids-pipeline output. Additionally, an `*_xcorr-artifact.npz` file stores part of the cross-correlation plot. A helper script combines all `.json` files to generate an overview over every recording as `.xlsx` [maybe update to store as .csv]

Cross correlation plot

The cross-correlation is computed between the horizontal EOG (HEOG) signal and the eye-tracker's horizontal gaze position ("L POR X [px]" and "R POR X [px]", averaged if both are present). If both streams are correctly time-aligned, EOG potential and gaze position movement should co-occur and produce a strong correlation at (or very near) zero lag.

[very unsure about this section. jevri said these plots should look like sharp normal distributions, not like pyramids]

An exemplary cross correlation plot between gaze position and EOG can be found in the documentation of the EYE-EEG toolbox. As can be seen, the cross correlation peaks at zero lag with a parabola-like shape at the tip, and with symmetric linear lobes.

[when i was identifying EOG channels, jevri said it should have step/box-like shapes in the signal. also, the gaze position signal (L POR X) also has box-like shapes. it would make sense if xcorr between two step functions gives a triangle with linear lobes. i need actual pictures of the data if i want to make this claim though]

[@jevri: do you have an idea why this xcorr curve could look so different from the one outputted by the mne-bids-pipeline fork? I looked at the code and this should also just be the xcorr between EOG and gaze position (not gaze velocity or something like that)]
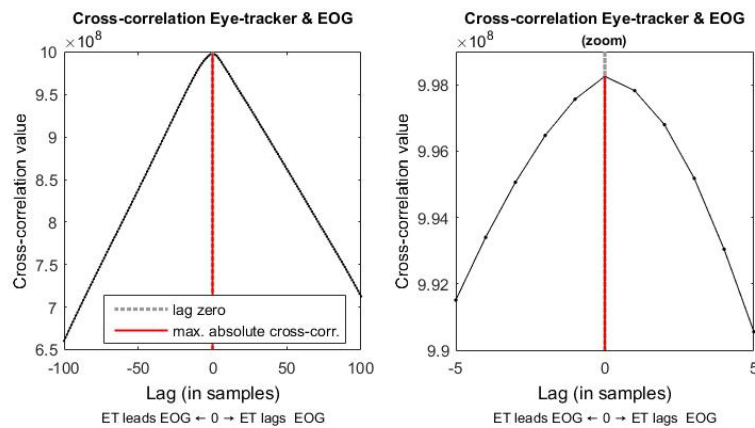
Figure 3.1: Cross-correlation plot from EYE-EEG [6]

[maybe mention band pass filtering done in the fork]

To automatically assess the desirability of the cross correlation curve, some dedicated measurements are made, including:

- **Peak position and height:** The peak position should be close to 0 lag. The height strongly depends on the subject, but it should be the highest peak in the entire plot. Additionally, measurements of the second highest peak in the plot are made, so it can be determined how prominent/distinct the center spike is. If the second highest peak of the cross correlation curve is very far from 0 lag and almost as high, it may be an indicator that the center peak is just coincidence [reformulate, very bad paragraph]

- **Similarity to 'template curve':** As seen in Figure 3.1, an optimal cross correlation curve will have linear lobes downwards from its peak. Using the value of the plot at zero lag and computed endpoints for the lobes on either side, a triangular 'template curve' is constructed, that is expected to have high similarity with the cross correlation curve iff the signals are well aligned [reformulate and also elaborate here. specifically how exactly are the endpoints for the lobes chosen: usually where the xcorr plot crosses y=0, but if this does not happen, it attempts to find prominent valleys instead]. [the similarity with the template curve is calculated in different ways like cosine similarity and kullback leibler divergence]

- [lobe steepness of the template curve: steeper is better. also symmetric (i.e. similar steepnesses for both lobes) is probably also better?]

- [kurtosis of the cross correlation curve. todo]

- [how many peaks are around the center of the plot? to see if it is 1 prominent peak or many peaks]

[could it make sense to test the xcorr curve with different offsets and see how high the highest peak is with these random offsets? if the peaks are just as high as when synchronized, the xcorr curve is probably worthless for this subject]

3.1.6 Evaluation of quality metrics

[add histograms of select metrics where interesting]
[create scatterplot of median_abs_sync_error_ms vs snr. could be interesting to see how strong the correlation is, and whether there are any clusters]
[research if there is an established treshold for xcorr value in literate]
[perhaps the only 'objective' metric of synchronization quality is the median_abs_sync_error_ms. if it's off by more than 2 ET samples on average, synchronization is probably off]
[multiple things have already been attempted here, but probably nothing that will remain in its current form, so not documenting it here yet]

3.2 Preprocessing

[entire section is just notes for now]

```
# needs to be set to ignore analysis and only do preprocessing
task_is_rest = True
```

```
# these need to be set also
# see https://github.com/mne-tools/mne-bids-pipeline/issues/1020
rest_epochs_duration = 5
rest_epochs_overlap = 0

# these are forced by ICALABEL
l_freq: float | None = 1
h_freq: float | None = 100
eeg_reference = "average"

# data was recorded in the US
notch_freq = 60

# size of the dataset requires icalabel
ica_algorithm = "picard-extended_infomax"
ica_use_icalabel = True
ica_reject = "autoreject_local"

# no epoch rejection, we need continuous data
reject = None

# enable eye-tracking synchronization
sync_eyelink = True
sync_eye = True

# cover all symbol search trigger events
sync_eventtype_regex     = r"(?:trialResponse|newPage)" #r"trialResponse"
sync_eventtype_regex_et  = r"# Message: (?:14|20)" #r"# Message: 14"

# HBN setup paper lists the following electrodes as being EOG electrodes:
# 8, 14, 17, 21, 25, 125, 126, 127, 128
# looking at the 3d montage, this is the only
# set of these electrodes that makes sense here
eeg_bipolar_channels = {
    "HEOG": ("E8", "E25"),   # left vs right outer canthus
    "VEOG": ("E21", "E17"),  # nasion vs left inner canthus
}

eog_channels = ["HEOG", "VEOG"]
sync_heog_ch = "HEOG"
sync_et_ch = ("L POR X [px]", "R POR X [px]")
```

```
montage = mne.channels.make_standard_montage("GSN-HydroCel-128")
eeg_template_montage = montage
drop_channels = ["Cz"]
```

4 TODO

- use consistent tense when writing (currently switching between past and present)
- properly cite everything

Bibliography

[1] Neuroelectromagnetic Data Archive and Tools Resource (NeMAR). *NeMAR Data Explorer: HBN-EEG*. n.d. URL: https://nemar.org/dataexplorer/local?search=HBN-EEG (visited on 01/07/2026).

[2] *International Neuroimaging Data-Sharing Initiative (INDI) - Registry of Open Data on AWS*. URL: https://registry.opendata.aws/fcp-indi/.

[3] *Index of /indi/cmi_healthy_brain_network/release*. URL: https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/release/.

[4] *The BIDS Validator - BIDS Validator documentation*. URL: https://bids-validator.readthedocs.io/en/latest/.

[5] Lindsay M. Alexander et al. "An open resource for transdiagnostic research in pediatric mental health and learning disorders". In: *Scientific Data* 4.1 (Dec. 2017). ISSN: 2052-4463. DOI: 10.1038/sdata.2017.181. URL: http://dx.doi.org/10.1038/sdata.2017.181.

[6] *The BIDS Validator - BIDS Validator documentation*. URL: https://www.eyetracking-eeg.org/tobii_eeg.html.

All links were last followed on October 5, 2020.

A My first appendix

Sample Appendix

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift