A content segmentation and scrolling display project based on Chinese language corpus

Introduction

This project aims to conduct a novel reading task with **real-time highlighted character** based on **Chinese language corpora** while simultaneously recording participants' brainwave (EEG) and eye movement (eye-tracking) data. The collected data can be utilized for EEG language decoding and related research endeavors.

Device Models

EGI: 128-channel

eyetracker: Tobii Glass 3

Environment

This project must be based on Python version 3.10!!! You can use Anaconda to create a new environment to run this project. The command is as follows:

```
conda create -n psychopy_tobii python=3.10
```

Then you can activate this environment to install packages we need:

```
conda activate psychopy_tobii
```

Psychopy

The entire experimental program is written based on Psychopy. You can download Psychopy either through the command line using the following command or by downloading it from the Psychopy official website: https://www.psychopy.org/.

```
pip install psychopy
```

There are two modes in Psychopy: builder and coder. In our project, we use the coder to implement our experiment.

g3pylib

This project utilizes g3pylib, a Python client library for tobii Glasses3. You can install this package by following the instructions here: https://github.com/tobiipro/g3pylib

You should first clone the package:

```
git clone https://github.com/tobiipro/g3pylib.git
```

Then you can install the package under the path of the cloned package using the following command:

```
pip install .
```

egi-pynetstation

egi-pynetstation is a python package to enable the use of a Python-based API for EGI's NetStation EEG amplifier interface. Install it using the following command:

```
pip install egi-pynetstation
```

For more information about this package, you can visit this website: https://github.com/nimh-sfim/egi-pynetstation

Code Explanation

The project mainly consists of two parts: you should first run `cut_Chinese_novel.py` to process your `.txt` formated Chinese novel for sentence segmentation, obtaining the corresponding formatted `.xlsx` files. Then you can run `PlayNovel.py` to start the main experiment.

Sentence Segmentation

`cut_Chinese_novel.py` is used to process the novel into a suitable format for presentation in the Psychopy program. The novel content will be segmented into a series of units, each containing no more than 10 Chinese characters, and adjacent three units will be combined to form the content of each frame during the screen presentation. The same combination of adjacent units will be repeated multiple times to facilitate the subsequent highlighting display of each individual character. Additionally, you can specify to divide the novel from certain chapters into several parts and output them as separate files.

The input for this script must be a Chinese novel content in `.txt` format. The format of the novel content should meet the following requirements:

1. **The beginning of the file should contain the preface section with a sign Ch0**. The preface will be used as a simulation before reading the formal chapters.

2. Each subsequent chapter should start with a sign **Ch + <chapter number>**.

Below is an example:

```
Chinese_novel.txt
Ch0
This is the preface of the novel
Ch1
Chapter 1 of the novel
Ch2
Chapter 2 of the novel
...
...
...
```

In real experiments, considering issues like eye tracker battery life, we must perform battery replacement during the experiment. Due to technical constraints, we have to temporarily stop the experiment program during battery replacement. Therefore, we may divide a complete novel into multiple parts and present each part during each complete program run. For example, we may divide a novel with 30 chapters to 4 parts, and we will run the program several times to finish the entire novel content.

In this script, you can specify the chapters you want to divide the novel by changing parameter `divide_num`, and the program will divide the chapters and output the files containing these different parts.

The output files are as follows:

`segmented_Chinese_novel_preface.xlsx`: The processed preface of the novel.

`segmented_Chinese_novel_main_1.xlsx` : The processed main body of the novel content.

`segmented_Chinese_novel_main_2.xlsx` : The processed main body of the novel content.

...

(number of files depends on your setting)

We provide a `.txt` file of the novel *The Little Prince* and the 5 processed `.xlsx` files as an example.

Parameters

Parameter	type	usage
divide_nums	str	Breakpoints which you want to divide your novel (comma-separated). e.g. If you want to break the novel from chapter 8, 16, 24, you should pass the argument as --divide_nums=8,16,24
Chinese_novel_path	str	Path to your `.txt` Chinese novel content

Main Experiment

Main Procedure

`PlayNovel.py` is used to run the main experiment based on Psychopy.

The program includes multiple modes, allowing you to choose whether to connect to the EGI and eye tracker for the experiment. The experimental procedure mainly consists of eye tracker calibration (if the eye tracker mode is selected), practice reading, formal reading, and breaks between chapters. During the reading process, the novel will be presented on the screen with three lines per page, and each line will not exceed ten Chinese characters (excluding punctuation marks). On each page, the middle line will be highlighted as the focal point, while the upper and lower lines will be displayed with reduced intensity as the background. Each character in the middle line will be sequentially highlighted for a certain duration, and participants will be asked to follow the highlighted cues to read the novel content. Various parameters, such as individual character highlight duration, highlight color, etc., can be adjusted by modifying the corresponding settings.

Parameters

Detailed explanations of the parameters will be provided below. Note that you must change the bold parameter to your own settings, or there may be some errors and calculation inaccuracies.

Parameter	type	default	usage
highlight_color	str	'red'	Highlight color of the characters
add_eyetracker	bool	True	Whether conneting to the eyetracker
add_mark	bool	True	Whether connecting to the egi device
shift_time	float	0.35	The shifting time of the highlighted character
host_IP	str	'10.10.10.42'	The IP address of the net station (The computer which runs this experiment)
egi_IP	str	'10.10.10.51'	The IP of the egi device
eyetracker_hostname	str	"TG03B-080202024891"	The serial number of the eyetracker
novel_path	str	"segmented_Chinese_novel_main.xlsx"	The path of the .xlsx format novel you want to play
preface_path	str	"segmented_Chinese_novel_preface.xlsx"	The path of the .xlsx format preface you want to play
fullscreen	bool	True	Whether to set a full screen
rest_period	int	3	The chapter interval of rest
force_rest_time	int	15	The forced rest time
distance_screen_eyetracker	float	73	distance from the center of the screen to the center of the eyetracker in centimeter
screen_width	float	54	The width of the screen
screen_height	float	30.375	The height of the screen
screen_width_height_ratio	float	16/9	The ratio of the screen width to screen height
eyetracker_width_degree	float	95	The horizontal scanning range of the eye-tracking camera in degree (both sides together)
eyetracker_height_degree	float	63	The vertical scanning range of the eye-tracking camera in degree (both sides together)
isFirstSession	bool	True	Whether this is the first session of the experiment, this will determine whether to display the preface before the formal experiment.

Notice: As mentioned in the previous section, we may run this script multiple times in the experiment to sequentially present each part of the novel. You need to specify the parameter "isFirstSession" every time you run this program to let it know if this is the first playback. If the value is "True," the program will play the preface to do practice reading before the formal reading begins. If it is "False," the practice reading part will be skipped, and the program will start directly from the main content.

EEG Markers

If you use the EGI device to record EEG signals during the experiment, our program will place markers at certain corresponding time points. These markers will assist you in aligning eye tracker recordings and EEG signals, as well as locating texts corresponding to specific segments of EEG signals.

The detailed information of markers are shown below:

```
EYES: Eyetracker starts to record
EYEE: Eyetracker stops recording
CALS: Eyetracker calibration starts
CALE: Eyetracker calibration stops
BEGN: EGI starts to record
STOP: EGI stops recording
CH01: Beginning of specific chapter (Numbers correspond with chapters)
ROWS: Beginning of a row
ROWE: End of a row
PRES: Beginning of the preface
PREE: End of the preface
```

Calibration Coordinate Transformation

In this experimental program, we designed a personalized calibration procedure. A dot will appear sequentially at the four corners and center of the screen, each staying for 5 seconds. Participants are required to fixate on the center of the dot to complete the calibration. For each dot, we record the middle and later segment of the participant's gaze data (from 3s to 4s) and calculate the average point of gaze as the participant's mean fixation point. We then compare the average fixation point with the actual center position of the dot to calculate the error. By averaging the errors from all five dots, we obtain the final calibration error. If the final error is below the predetermined error threshold, we consider the calibration as successful. If the calibration is not successful, the experimental program will automatically return to the calibration phase and repeat the process until calibration is achieved.

In order to align the coordinate systems of the eye tracker and the Psychopy program to obtain the actual positions of gaze points on the screen, we derived a transformation formula between the coordinate systems of the eye tracker and the Psychopy program using geometric relationships. This formula was then applied during the calibration process. The specific relationship is as follows:

$$x_{\text{eyetracker}} = \left(\frac{W \cdot x_{\text{psychopy}}}{d \cdot r \cdot \tan(\text{width\_degree}/2)} + 1 \right) \cdot \frac{1}{2}$$

$$y_{\text{eyetracker}} = \left(1 - \frac{H \cdot y_{\text{psychopy}}}{d \cdot \tan(\text{height\_degree}/2)} \right) \cdot \frac{1}{2}$$

w here :

$(x_{eyetracker}, y_{eyetracker})$ *is the coordinate in the eyetracker coordinate system*

$(x_{psychopy}, y_{psychopy})$ *is the coordinate in the psychopy coordinate system*

W : *the width of the screen,* H : *the height of the screen,* r : *the ratio of the width to the height*

width_degree : *the horizontal scanning range of the eyetracking camera in degree (both sides together)*

height_degree : *the vertical scanning range of the eyetracking camera in degree (both sides together)*

Data Acquisition and Processing

The data is saved in the order of recording with a structure and naming convention as follows:

```
.
├─sub_01
│   ├─eegData
│   │       sub_01_eeg_01.rar
│   │       sub_01_eeg_02.rar
│   │             ...
│   │             ...
│   │
│   └─eyetrackingData
│           sub_01_eyetracking_01.rar
│           sub_01_eyetracking_02.rar
│                 ...
│                 ...
│
├─sub_02
│   ├─eegData
│   │       sub_02_eeg_01.rar
│   │       sub_02_eeg_02.rar
│   │             ...
│   │             ...
│   │
│   └─eyetrackingData
│           sub_02_eyetracking_01.rar
│           sub_02_eyetracking_02.rar
│                 ...
│                 ...
│
└─...
```

Experiment Procedure

Below are the operational steps and an example of starting the project from scratch, using the novel *The Little Prince* as an example.

Activate Environment

First, activate the environment we set up before, and then navigate to the directory where the project is located.

```
conda activate psychopy_tobii
cd <your_path_to_project>
```

Novel Segmentation

Take a `.txt` novel file that meets the format requirements (format requirements can be found in the "Sentence Segmentation" section of the Code Explanation) as input. Specify the parameter to divide the text into several parts. Run `cut_Chinese_novel.py`, and you will obtain the corresponding number of `.xlsx` files. Here, we divided the novel into 4 main parts, resulting in 5 files (4 main body parts and 1 preface part).

```
python cut_Chinese_novel.py --divide_num=8,16,24 --
Chinese_novel_path=xiaowangzi_main_text.txt
```

Main Experiment

- First, connect the EGI equipment and the eye tracker.

- Next, adjust the parameters and run the main program. Set the addresses for the main body and preface parts in the variables `novel_path` and `preface_path`, respectively. Adjust the parameters `add_mark` and `add_eyetracker` to decide whether to connect to the EGI and eye tracker. Change `host_IP`, `egi_IP`, and `eyetracker_hostname` to the IP numbers of your own devices. Set `isFirstSession` to True during the first run to include the preview session. Other adjustable parameters can be found in the "Parameters" section of the Code Explanation under Main Experiment. Note that you may need to modify some size and distance-related parameters according to your own setup. In subsequent runs, change the `novel_path` to read different parts of the novel and set `isFirstSession` to False.

```
python PlayNovel.py --add_mark --add_eyetracker  --preface_path=<your preface path> --
host_IP=<host IP> --egi_IP=<egi IP> --eyetracker_hostname=<eyetracker serial number> --
novel_path=<your novel path> --isFirstSession
```

Here is our own settings as an example:

```
First time:
python PlayNovel.py --add_mark --add_eyetracker  --
preface_path=segmented_Chinense_novel_preface.xlsx --host_IP=10.10.10.42 --
egi_IP=10.10.10.51 --eyetracker_hostname=TG03B-080202024891 --
novel_path=segmented_Chinense_novel_main_1.xlsx --isFirstSession

Second time:
python PlayNovel.py --add_mark --add_eyetracker  --
preface_path=segmented_Chinense_novel_preface.xlsx --host_IP=10.10.10.42 --
egi_IP=10.10.10.51 --eyetracker_hostname=TG03B-080202024891 --
novel_path=segmented_Chinense_novel_main_2.xlsx

...
```

- **During the forced break period, the EGI system will be disconnected.** *At this time, you need to restart the EGI system to ensure it is in a running state before the participant continues the experiment, or the program will crash!!!*

- **At the end of each experimental session, it is necessary to replenish the saline for the participant's EEG cap and replace the eye tracker's batteries to ensure sufficient power. Start and reconnect the eye tracker, and restart the EGI system.** *Remember not to disconnect the eye tracker or replace its batteries during the experiment (including the rest periods) as doing so may cause the program to crash!!!*

- The main process of the experiment includes: calibration - preface session (only in the first part) - formal reading - rest (including mandatory rest and participant-initiated rest periods). After each rest period, recalibration will be performed.

    - Calibration

      Note: When calibration fails multiple times, the experimenter can choose to skip the calibration and proceed directly to the reading section by pressing the right arrow key on the keyboard at the calibration failure prompt page.

    - Preface Reading

    - Formal Reading

    - Rest

      restart the EGI system to ensure it is in a running state before the participant continues the experiment