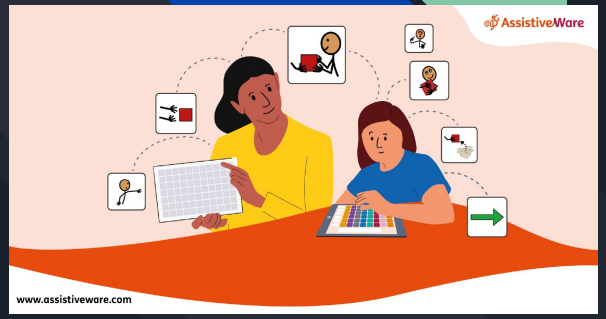
Final Project Advanced HCI

Michele Minniti

Well being digital education



- Inclusive learning
- Mental health issues
- Digital stress
- Augmentative and alternative communication (AAC)

We want to address the problem of augmentative learning for children in the autistic spectrum or with ADHD disturb

Literature review first delivery

Child play dataset

Emotion recognition

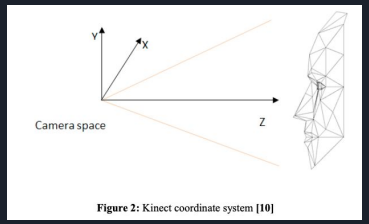
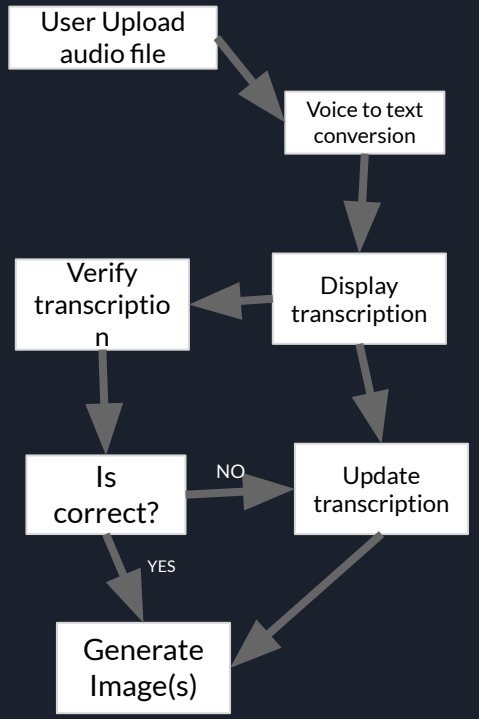


Figure 2: Kinect coordinate system [10]

Subject	MLP	3-NN
1	0.94	0.97
2	0.96	0.96
3	0.90	0.98
4	0.74	0.90
5	0.96	0.96
6	0.93	0.97
Average	**0.90**	**0.96**

Table 2: The results of the subject-dependent classification

Speech to text



Literature review second delivery

We started by analyzing several papers on the topic of the idea proposed and in particular regarding “ASD child learning and their needs”.

We analyzed the existing applications and technologies that support these kind of users and statistics about the learning process.

what resulted from the analysis of the papers is:

- Make products more portable
- Most existing products are only in the educational sector, and do not take into account the emotional and social sphere
- Some insightful papers of researchers who already applied an attention control mechanism has revealed an important workload to associate weights with given emotion and gaze responses

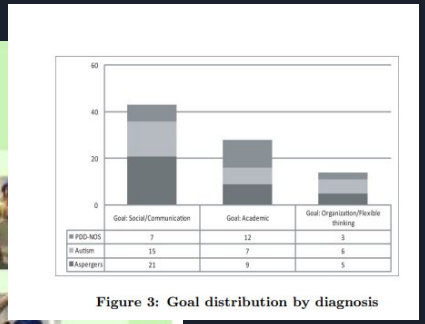


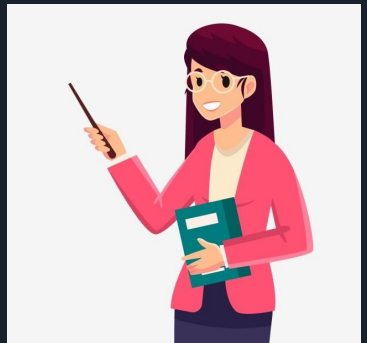
Figure 3: Goal distribution by diagnosis

Studio Assistant

- SPEECH TO IMAGE GENERATION: based on the first paper, using a diffusion model that convert speech audio in to a textual input through a text encoder; then the text will be used as input and guidance for the image generation, correlated to the lesson or teacher's explanation.

- CONCENTRATION AID: based on the second and third paper, we want to use ML to recognize emotion by a camera and gaze detection, with the aim of better understand the emotions, and understand when it could be distracted through gaze, and therefore help him to better understand the concepts, in relation to his state of mind and concentration.

- Multimodal studio assistant to help with AAC students with special needs
- Image generation through speech to visualize the subjects of the explanation
- Gaze control and emotion recognition to assess difficulties or give positive feedback with gamification setup
- Enhancing teaching-learning experience and overcome educational barriers

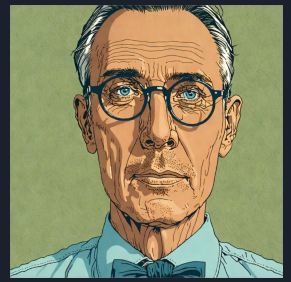


Personas

*Generated with Stable Diffusion

CHILD USER DESCRIPTION:

- **Name:** Rosie Ortiz
- **Age:** 7 y. o.
- **Characteristics:** Rosie is a young child with Autism Spectrum Disorder (ASD) who struggles with communication in her day-to-day activities. Rosie loves colorful visuals, somewhat repetitive patterns, and simple interfaces. She is sensitive to loud sounds, so she prefers quiet, calm settings.
- **Goals:** Learning at school in a comprehensible and fun way in order to not get bored.
- **Challenges:** Needs positive reinforcement to not get frustrated nor distracted. Gets frustrated if she fails to achieve her goal after a few trials.
- **Key stakeholders**: Parents, Teachers

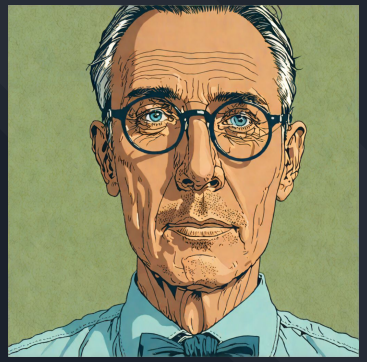


*Generated with Stable Diffusion

TEACHER USER DESCRIPTION:

- **Name:** Marc Reds
- **Age:** 65 y. o.
- **characteristics:** Marc is a support teacher specialized in ASD children. Has more than 30 years of experience in the field and has found interesting experimenting with new technologies to enhance his teaching methods.
- **Goals:** Support the child with ASD integrating innovative technological tools to improve communication and learning and monitor his progress.
- **Tech familiarity:** Simple but effective, knows how to run a software on his pc or an app for tablets.
- **Challenges:** Needs a comprehensive interface, variety of possibilities to improve communication with the child

Scenario
TEACHER USAGE



CONTEX:
Marc Reds, support teacher, uses the app to explain a school concept to a child with ASD, making the lesson more interactive thanks to the “interactive explanation” function. He can also monitor the emotion and concentration of the child with the emotion/distraction recognition.

STEPS:

1. Marc opens the app on the Rosie’s profile;
2. Marc select the “Interactive Explanation” mode;
3. Marc activates the “Distraction & Emotion recognition” mode;
4. Marc start the explanation. Example: “Today we learn what trees are. Trees have branches, leaves and roots!”;
5. The app generate in real-time image of a tree, branches, leaves and roots;
6. The app also provide a real-time feedback about the concentration and emotion of Rosie.
7. Marc consult the concentration/emotion feedback of rosie and continues explanation and talks about other topics, which are turned into pictures for Rosie;
8. Marc assign an exercise to Rosie about the topic explained before.
9. … MOVING TO SCENARIO OF STUDENT USAGE …

Scenario

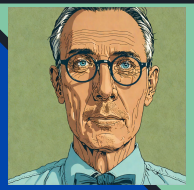
STUDENT USAGE

CONTEX:
Rosie uses the app to complete an elementary exercise, given by the teacher, while the app monitors his emotional state to detect any difficulties or distractions.

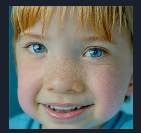
STEPS:

1. Rosie receives the exercise from teacher;
2. The app presents an instruction that explains what to do, and images to better understand the task;
3. Rosie start to complete the task
4. The functionality “Distraction & Emotion recognition” previously activated by teacher is running and is analyzing the gaze of rosie for distraction recognition and the facial expression to monitor her emotion and feelings during the exercise;
5. If Rosie gets DISTRACTED, the app emits a signal, like a colored pop-up appear. For example: a cartoon character telling her to continue the exercise.
6. In case Rosie becomes SAD/NERVOUS/AGITATED the app emits a signal like a colored pop-up appear. For example: a cartoon character telling her to relax, maybe take a break, and retry.
7. In the end, the app provides the outcome of the exercise (right/wrong) and the report on the distraction and the emotions felt by Rosie → FOR THE TEACHER

Detailed functionalities → “Interactive Explanation”



“albero”



What it does:

- Converts teacher's spoken instructions or explanations into vivid, illustrative images in real-time.

Purpose:

- Enhances comprehension by providing visual support aligned with verbal instructions.
- Bridges the gap between abstract concepts and understanding through visualization.

How it works:

- Speech recognition identifies keywords and contexts from the teacher’s speech.
- AI generates images dynamically based on detected inputs.

Example:

- Teacher says: *"Imagine a forest with tall trees and a clear blue sky."*
  - The software instantly generates an image of a serene forest.

“EMOTION RECOGNITION”

What it does:

- Detects emotional states in real-time through facial expressions.

Purpose:

- Helps teachers adapt their explanations based on the child’s emotional responses.
- Early identification of struggles or disengagement.

How it works:

- Uses AI models trained on diverse emotional datasets.
- Provides real-time feedback to the teacher on the child’s emotional state.

Key Benefits:

- Promotes emotional awareness.
- Promotes empathetic teaching approaches.

“GAZE RECOGNITION”

What it does:

- Tracks the child’s gaze.

Purpose:

- Alerts teachers if the child is losing attention or focus.

How it works:

- Analyzes gaze direction and duration using a webcam

Applications:

- Helps redirect focus through interactive activities.
- Teacher can adapt lesson content dynamically to REgain attention.

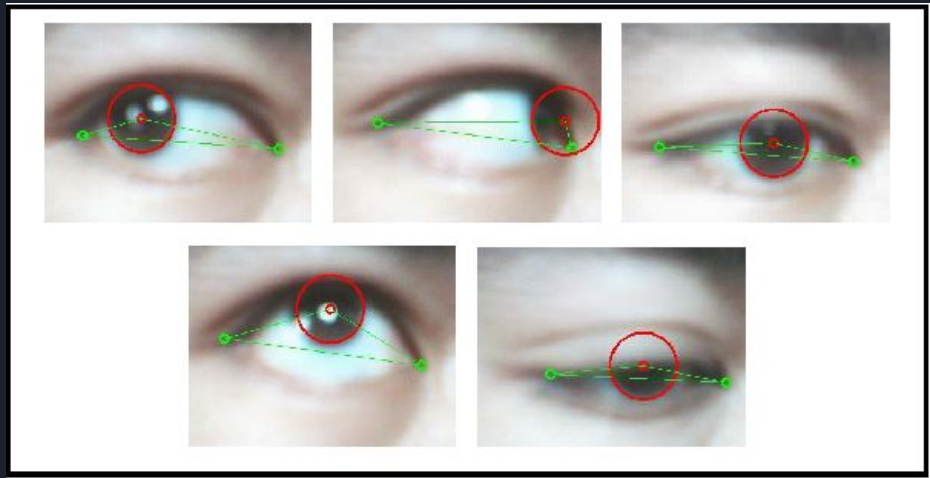


Figure 12. Results of the first experimental

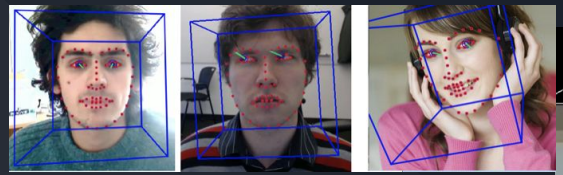
Synergy of Emotion and Gaze Recognition

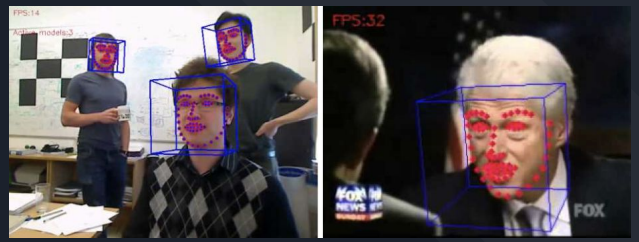
Combined Functionality:

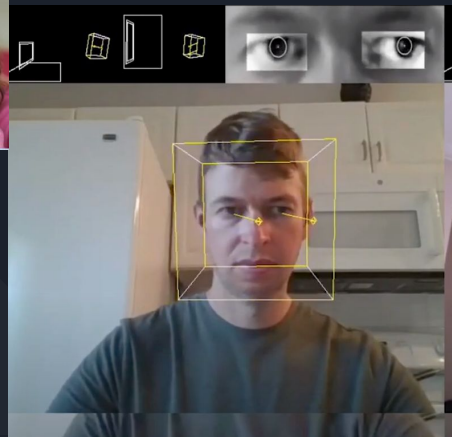
- Emotion and gaze data are combined for a better understanding of child's engagement.
- Example:
  - Gaze indicates focus, but emotion shows frustration → may be potential struggle with content.
  - Gaze and emotion indicate disengagement → prompts re-engagement strategies.

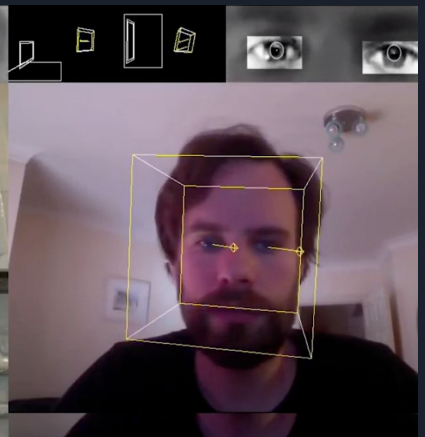
Outcome:

- Provides important feedback to teachers for adjusting lessons.
- Improves the overall learning experience.





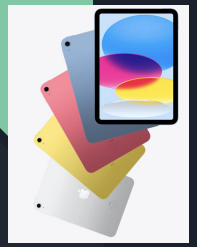


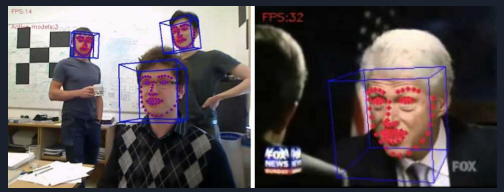


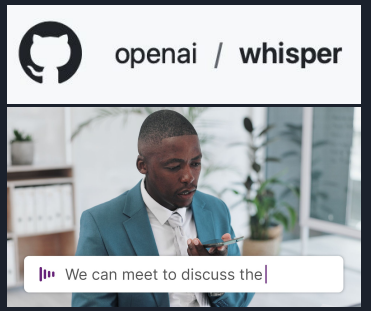
Hardware and software tools

TO DEVELOP OUR IDEA OUR IDEA

which basically will be an application







AS HARDWARE TOOLS:

1. LAPTOP
2. TABLET

AS SOFTWARE TOOLS:

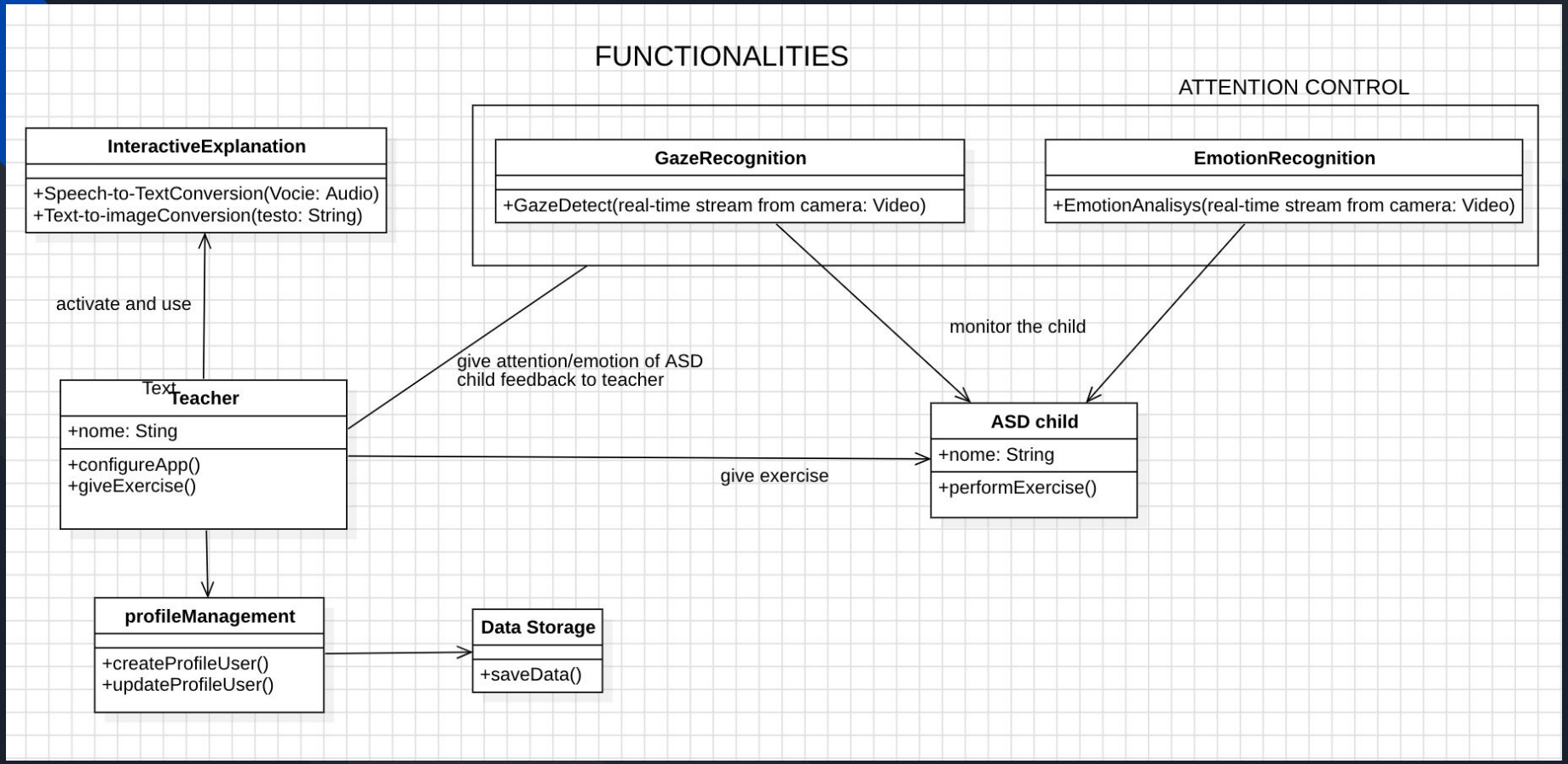
1. Facial and gaze detection tool
2. OpenAI whisper tiny version
3. Stable Diffusion for image generation

BACKEND

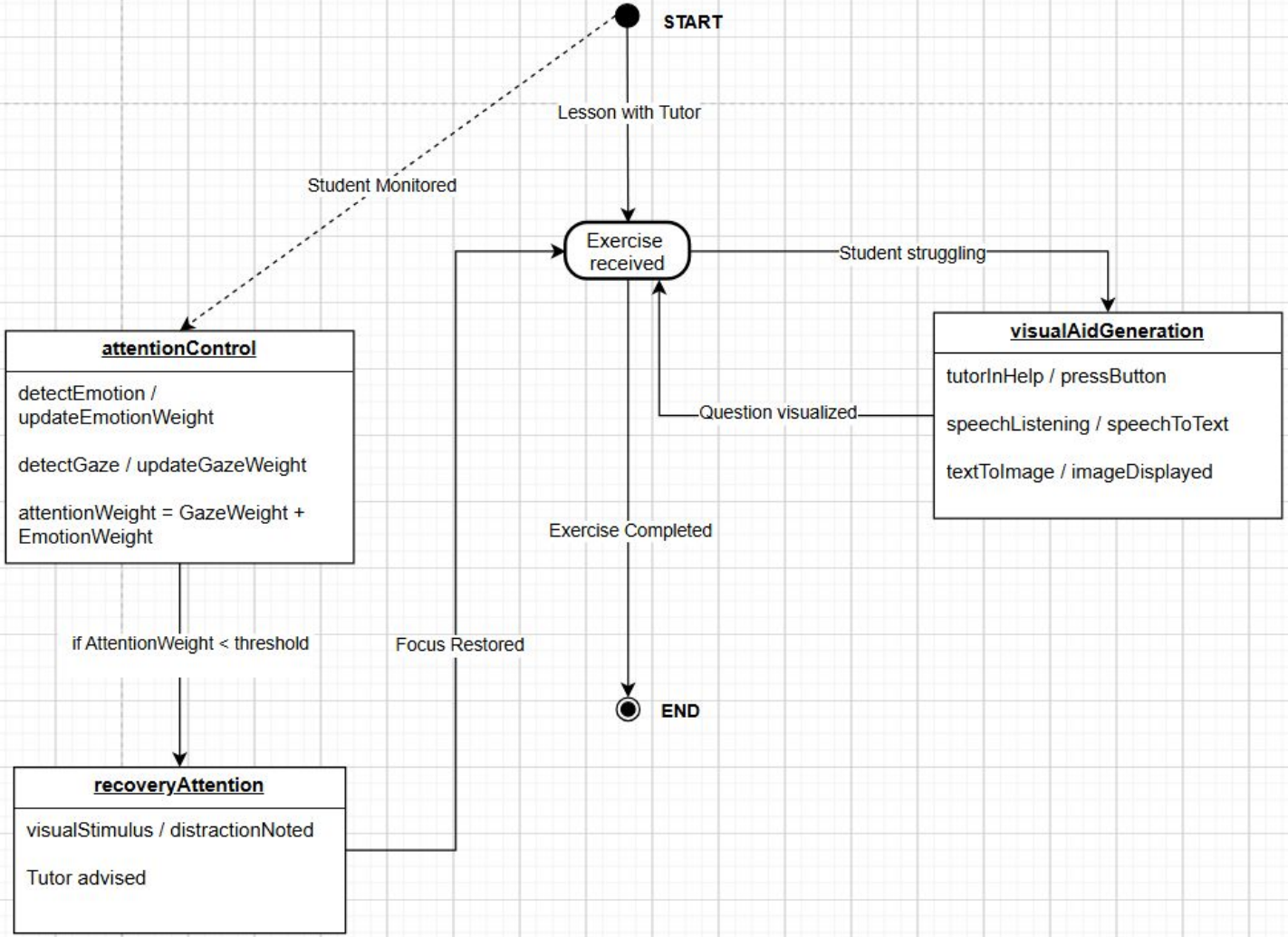
- Backend runned with flask
- Code written in python
- Dataset “CommonSenseQuestions” used to generate exercise
- Audio to image process handled by whisper tiny and SDXL Turbo
- Frontend handled with React

```
OPEN "questions" AS file
    READ content INTO QUESTIONS
CLOSE file

METHOD "GETQUESTION":
    SELECT random question FROM QUESTIONS
    RETURN JSON containing:
        - question text
        - multiple-choice options
        - correct answer

METHOD "GETVIDEO":
    STREAM video frames FROM frameSource()

METHOD "POSTAUDIO":
    RECEIVE audio file FROM request
    PROCESS file WITH whisper speech-to-text model
    STORE transcription result IN "text"

    RECEIVE JSON request containing "prompt"
    GENERATE image USING SDXL TURBO
```

UML COMPONENT MODEL



UML STATE CHART DIAGRAM



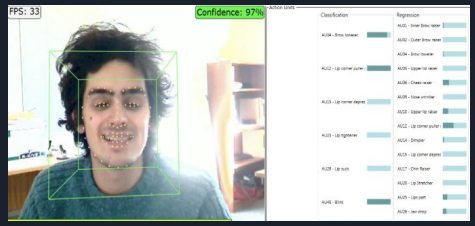
MULTIMODAL ANALYSIS

Emotion recognition and Gaze Detection

Feature extraction

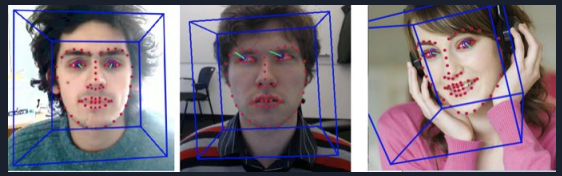
- **FACIAL LANDMARKS DETECTION**:

Identifies key points on the human face (such as eyes, nose, mouth…ecc) using computer vision and machine learning techniques. Facial landmarks contains crucial information for our analysis



- **EMOTION RECOGNITION**:

Analyzing and recognizes facial action units (AUs), we can track specific facial muscle movements. Similarity in AU detection can provide feedback about the success of face detection and alignment

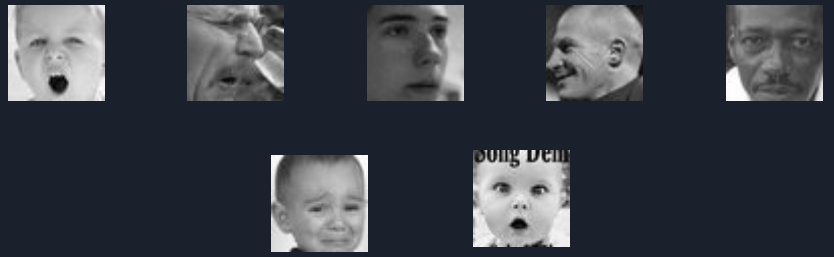


- **GAZE FEATURES**:

Determines the direction of a person’s gaze by analyzing eye position and corneal reflections. We can extract information about gaze direction and angle of pupil and blinking ratio to have information about the concentration of the subject.

Let’s test some methods

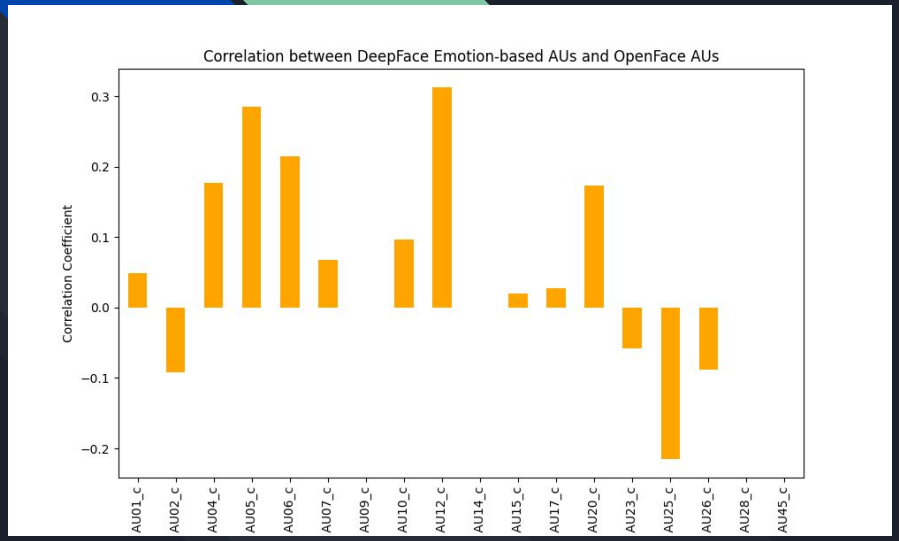
Here we have the results of various emotion recognition frameworks on the ananthu017 dataset (poor photo quality, non standard settings of lightning head pose etc.)



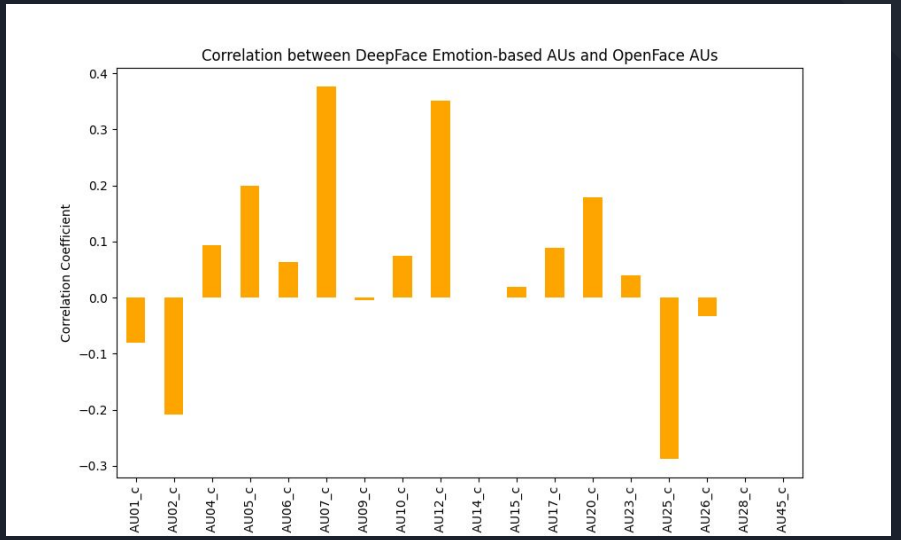
Opencv Accuracy: 55.14%
Retinaface Accuracy: 50.47%
MTCNN Accuracy: 52.93%
SSD Accuracy: 52.54%
DLIB Accuracy: 49.96%
Mediapipe Accuracy: 45.78%
Yolov8 Accuracy: 50.75%
Centerface Accuracy: 56.32%
Skip Accuracy: 56.46%

OpenFace and Gaze Tracking Accuracy to various test is around 60%

FACIAL DETECTION WITH VARIOUS TECHNIQUES

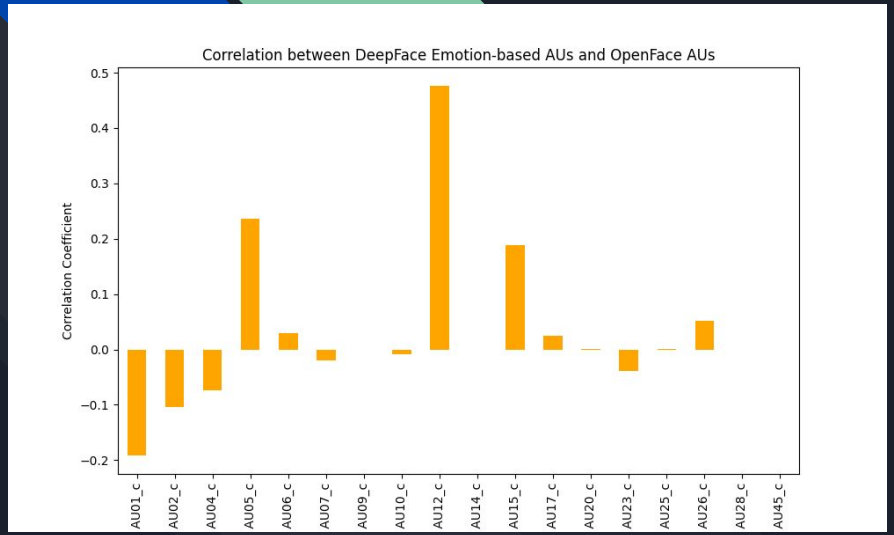


Openface vs opencv deepface implementation

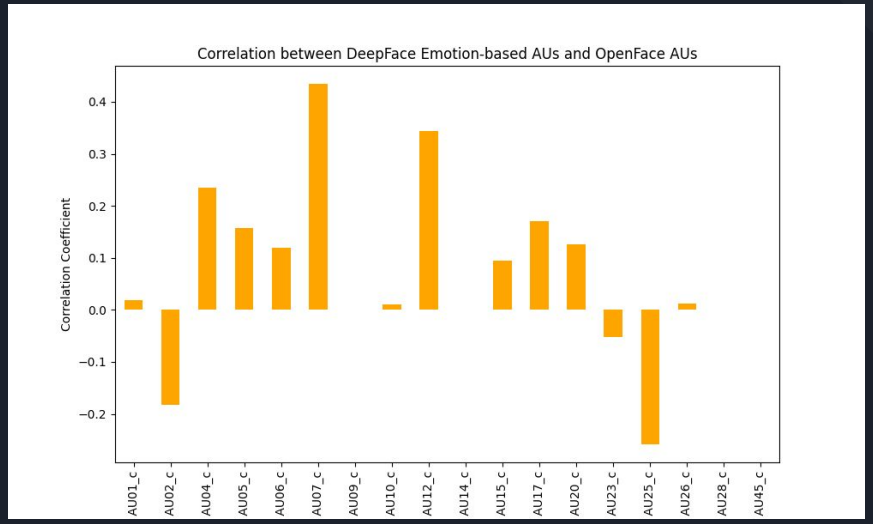


Openface vs yolov8 deepface implementation

FACIAL DETECTION WITH VARIOUS TECHNIQUES



Openface vs dlib deepface implementation



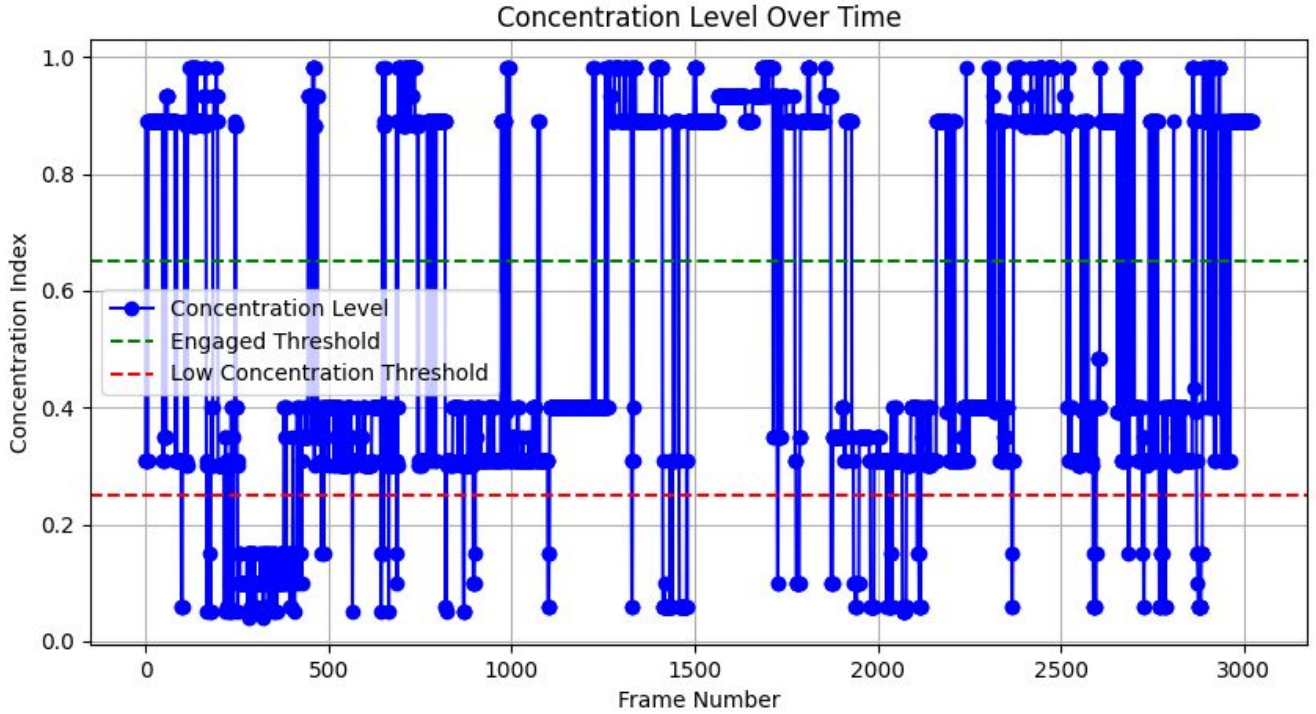
Openface vs mtcnn deepface implementation

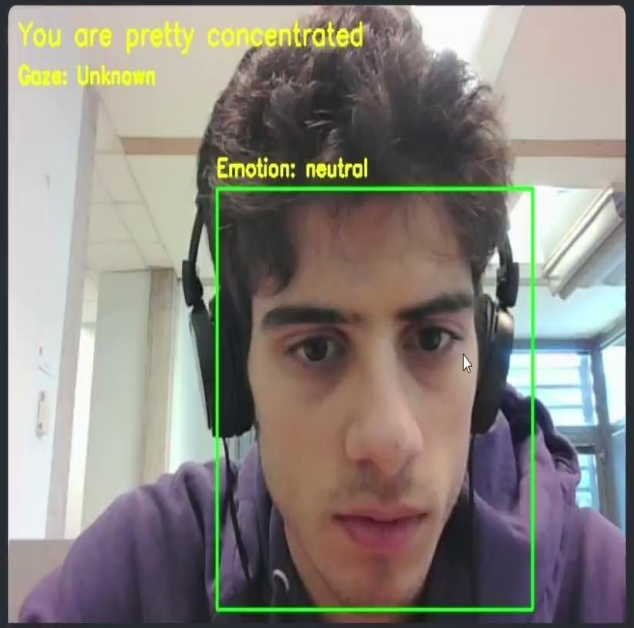
ATTENTION CONTROL

- Attention control based on emotion recognition and gaze analysis
- DeepFace and GazeTracking are used
- To every emotion recognized by DeepFace is associated a weight
- The gaze weight is computed analyzing eye openness (from euclidean distance by landmark points) and gaze direction
- Openface has been used as a baseline for the analysis.

```
FUNCTION calculate_eye_openess(eye):
    IF eye IS NULL:
        RETURN 0
    ENDIF

    COMPUTE A, B, C = euclidean_distance(vertical and
horizontal landmarks)

    COMPUTE EAR = (A + B) / (2 * C)
    RETURN EAR

FUNCTION concentration_index(emotion, gaze_weights):
    SET emotion_score TO emotion_weights[emotion]
    SET gaze_weights TO gaze_analysis(frame)
    COMPUTE concentration WITH emotion_score and
gaze_weights
    RETURN concentration

FUNCTION gaze_analysis(frame):
    PROCESS frame WITH gaze tracking(frame)
    COMPUTE openness = calculate_eye_openess(eye)
```

CONCENTRATION LEVEL





Exercise

What is a salesman responsible to do at work?

- A. traveling to chicago
- B. get fired
- C. books
- D. sell products
- E. service account

Submit Answer

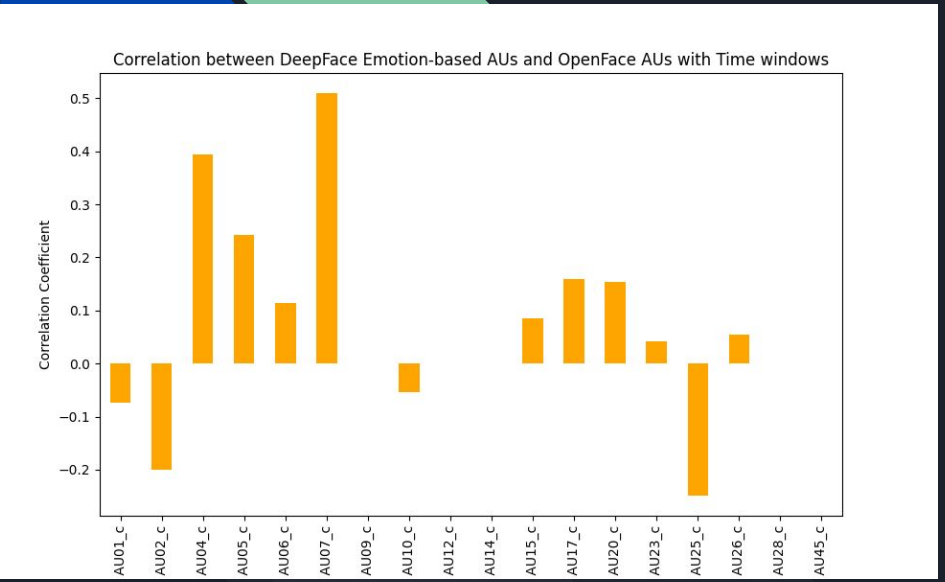
Record Audio

Start Recording

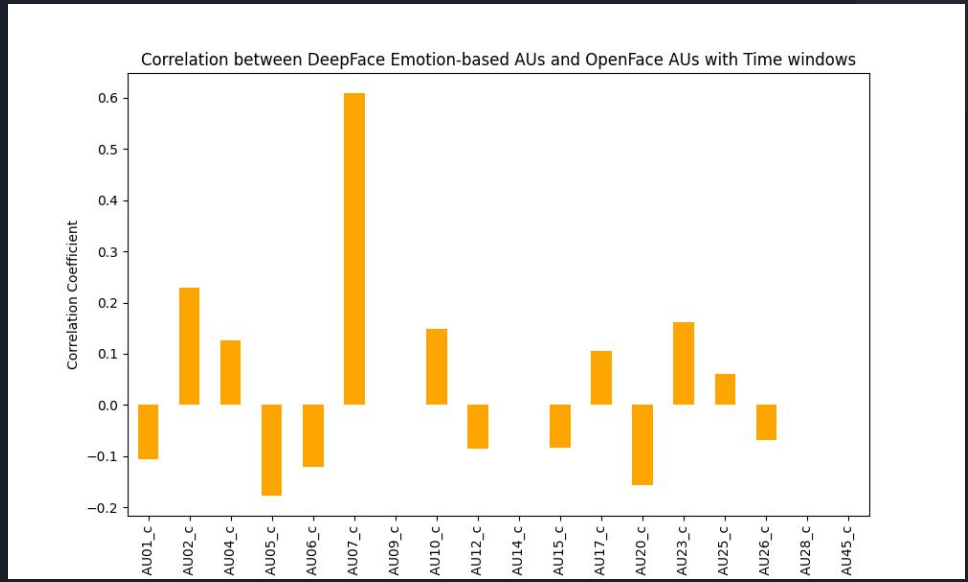
Generate Image

- Generate image from question

TIME WINDOW ANALYSIS

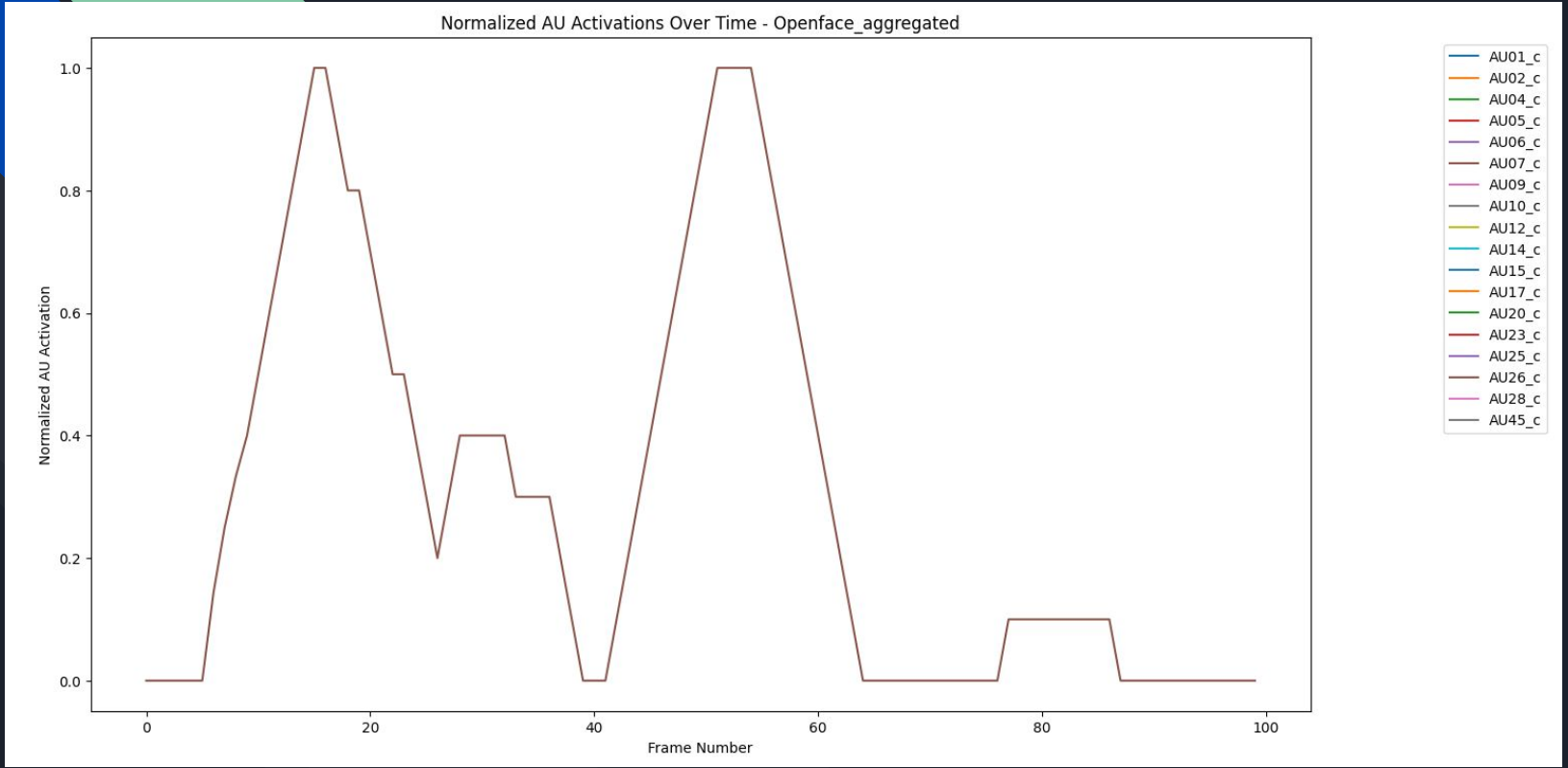


Openface vs mtcnn deepface implementation

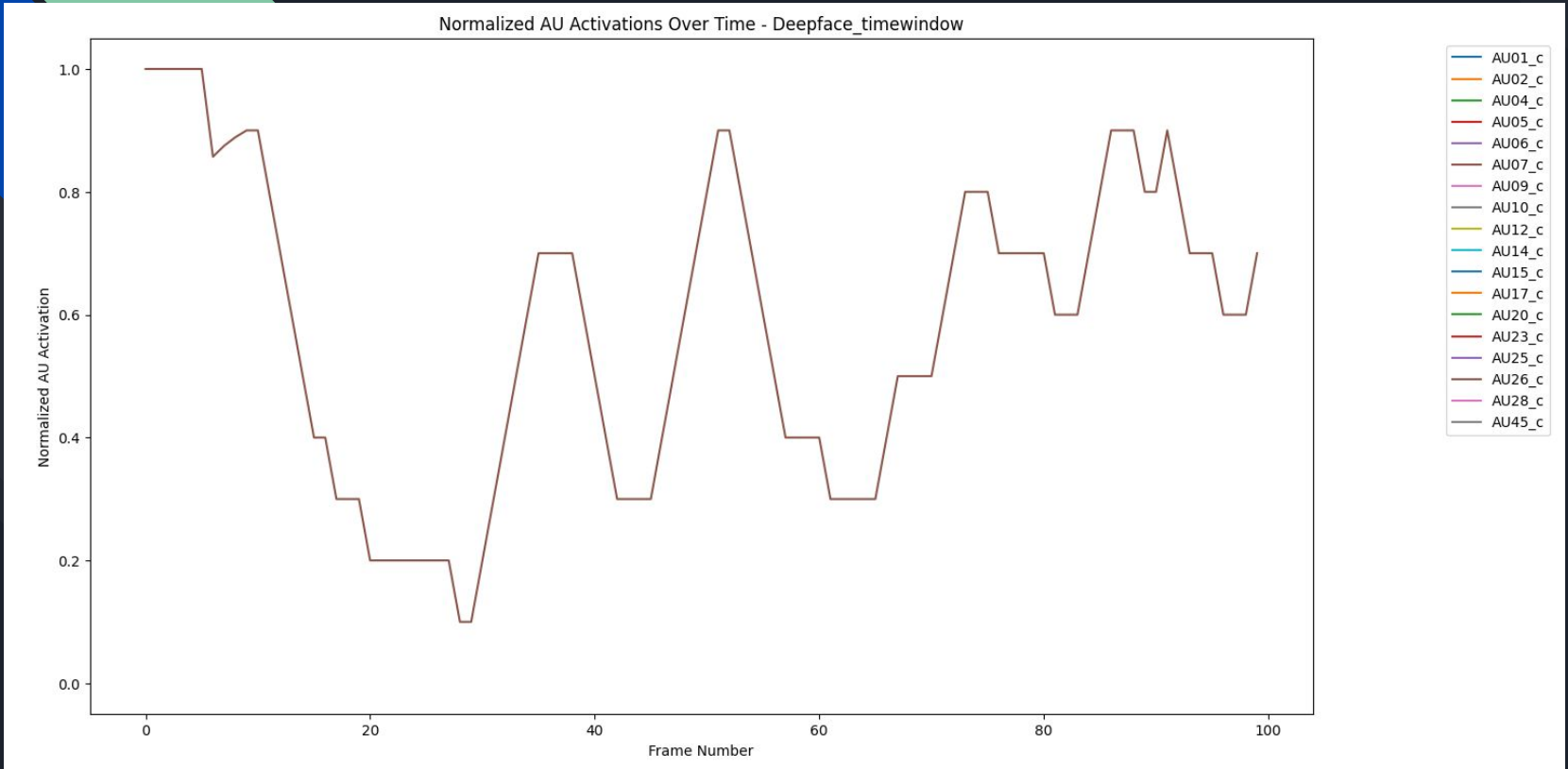


Openface vs centerface deepface implementation

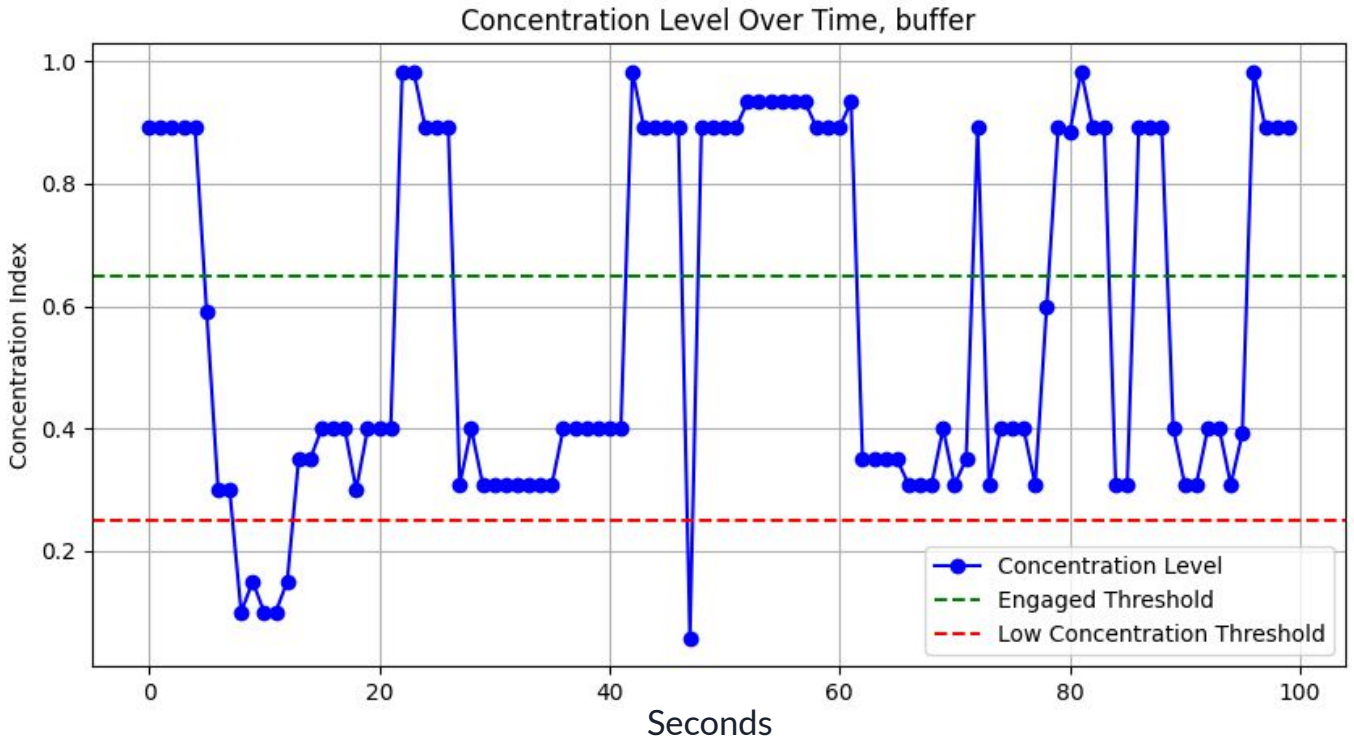
AU 26 OVER TIME OPENFACE



AU 26 OVER TIME DEEPFACE - SKIP



NEW RESULTS


Concentration Level Over Time, buffer
Concentration Index
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Seconds
Concentration Level
Engaged Threshold
Low Concentration Threshold

Exercise

Where can you go to use a piano in your neighborhood if you don't have one?

- A. music school
- B. music store
- C. neighbor's house
- D. lunch
- E. drawing room

Submit Answer

Record Audio

Start Recording

Generate Image

- Generate image from question

CONCLUSIONS

- Time window analysis (1-2 seconds with 50% Overlap) is more efficient to track feature in real time despite maintaining a good degree of accuracy
- AU and Deep Learning consider different features in their analysis, Deep learning method is more suitable for application in which the facial features can't be extracted in an ideal way
- Gaze Tracking and Emotion recognition combined can give us enough information to build a good attention control method using only a webcam as sensor
- Between machine learning methods concentration results there is a satisfying degree of similarity
- The attention control can work in real time with a satisfying degree of accuracy

References

1. *Akamsha Timande, Pallavi Borse, Vaishnavi Lande, & A.G.Sharma. (2024). Speech to Image Generation by Stable Diffusion Model. SSGM Journal of Science and Engineering, 2(1), 89–91. Retrieved from https://ssgmjournal.in/index.php/ssgm/article/view/116*
2. *Boluk, Nursena ; Kose, Hatice. / **Evaluating Gaze Detection for Children with Autism Using the ChildPlay-R Dataset**. 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition, FG 2024. Institute of Electrical and Electronics Engineers Inc., 2024. (2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition, FG 2024).*
3. *Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz J. Rak, Emotion recognition using facial expressions, Procedia Computer Science Volume 108, 2017. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050917305264*
4. Iacono, T., Trembath, D., & Erickson, S. (2016). The role of augmentative and alternative communication for children with autism: current status and future trends. *Neuropsychiatric disease and treatment*, 2349-2361.
5. Putnam, C., & Chong, L. (2008, October). Software and technologies designed for people with autism: what do users want?. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility* (pp. 3-10).
6. Khaleel, F. L., Ashaari, N. S., & Wook, T. S. M. T. (2020). The impact of gamification on students learning engagement. *International Journal of Electrical and Computer Engineering*, *10*(5), 4965-4972.
7. Bondioli, M., Buzzi, M. C., Buzzi, M., Chessa, S., Jaccheri, L., Senette, C., & Pelagatti, S. (2024). Guidelines for research and design of software for children with ASD in e-health. *Universal Access in the Information Society*, *23*(4), 1909-1930.
8. Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.
9. Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2025). Adversarial diffusion distillation. In *European Conference on Computer Vision* (pp. 87-103). Springer, Cham.
10. Cakmak, M. C., Okeke, O., Spann, B., & Agarwal, N. (2023, May). Adopting parallel processing for rapid generation of transcripts in multimedia-rich online information environment. In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (pp. 832-837). IEEE.
11. Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., ... & de Jesus Filipe, V. M. (2022, August). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *International Conference on Technology and Innovation in Learning, Teaching and Education* (pp. 52-68). Cham: Springer Nature Switzerland.