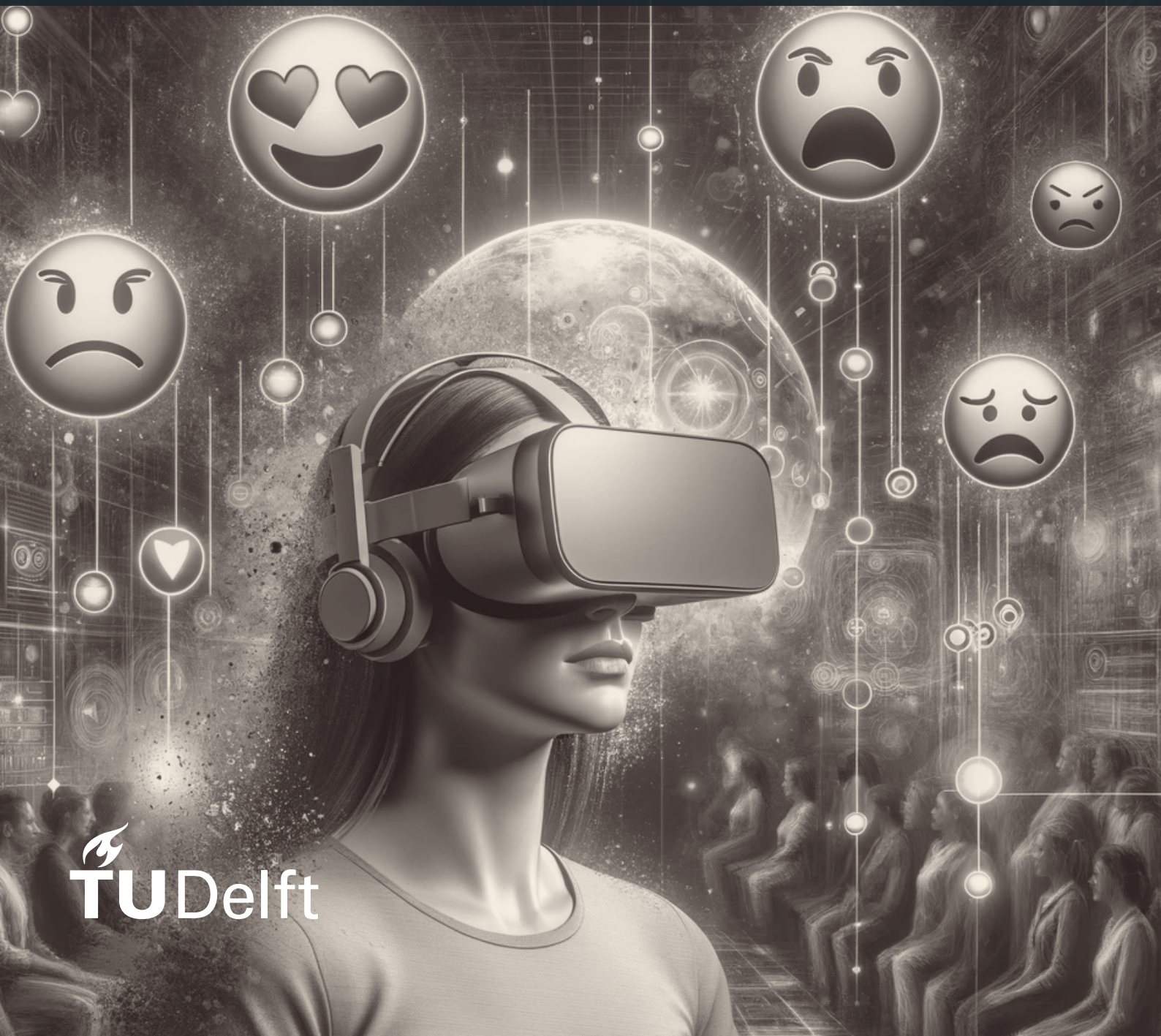
Emotion Recognition in Virtual Reality

Creation and validation of a VR-based multi-modal emotion recognition dataset

Bishwas Regmi



TU Delft

Emotion Recognition in Virtual Reality

Creation and validation of a VR-based multi-modal emotion recognition dataset

Thesis report

by

Bishwas Regmi

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on April 19, 2024 at 15:15

Thesis committee:

Chair:	Prof. dr. K.G. (Koen) Langendoen
Supervisor:	Dr. Guohao Lan
External examiner:	Dr. Xucong Zhang
Place:	Faculty of EEMCS, Delft
Project Duration:	April, 2023 - April, 2024
Student number:	4467655

Faculty of EEMCS · Delft University of Technology

Abstract

Emotion recognition in Virtual Reality(VR) has the potential to offer numerous benefits across various sectors such as mental healthcare, education, marketing, entertainment, etc. Although emotion recognition itself is a mature field, the sub-field of VR-based emotion recognition is still in its early stages of development. It was found that a limiting factor in the progress of this field is a lack of sufficient data for research and development of advanced deep learning models. Also, the equipment currently used to measure emotion-related signals is expensive and impractical for general usage. This thesis aims to support the progress in this field by creating a VR-based emotion recognition dataset using VR equipment only. This addresses the problem of insufficient data available for research and development, and also reduces the reliance on expensive and impractical equipment for emotion recognition.

To create a good quality dataset, several important things had to be addressed. First of all, the stimuli to evoke the emotions had to be carefully selected to ensure that genuine emotional responses are evoked and recorded in the dataset. Then, an efficient data collection system had to be created to ensure that the data collection process runs effectively, smoothly and consistently. Then, a proper labeling process had to be designed to annotate the data as accurately as possible. Finally, the compiled dataset was validated by showing that the chosen stimuli were effective in evoking the intended emotions. This was verified through the analysis of pupil response data, which is one of the recorded data modalities.

Preface

> "Your intellect may be confused, but your emotions will never lie to you."
>
> — Roger Ebert

I have been fascinated by the working of human emotions for a while and was very grateful to get the opportunity to do this project. I had hoped that I would learn to interpret my own emotions better through this project, which I certainly have but I have also come to realize that intellect may never be able to grasp the true depth and nuance of emotions. Perhaps this is not the purpose of intellect anyway. Spiritual masters since ancient times have emphasized the importance of quieting the intellect and feeling the emotions without any thoughts to get a full understanding of oneself. So it may not be unreasonable to assume that we are born with the natural capability to understand our emotions, and the intellect has a different purpose than interpreting emotions. Perhaps we always have a deeper understanding of our emotions and it is the function of the intellect to manifest this already existing understanding into speech and actions. We know instinctively when things are not as they should be, but our intellect usually fails to recognize the proper solutions. We are driven by our emotions but are often confused and lost because the intellect is not fully functioning as we would like it to. It is limited by knowledge, time and mental energy, and is prone to false beliefs and conditioning. Even if it seemingly has found a solution, we quickly realize that it is not the true solution that the emotions have been indicating so it goes on trying new things or repeats the same cycle with slight variations. Since machines are an extension of the intellect, perhaps they can assist the intellect to align with the intentions set by the emotions more effectively. This is my hope at least.

Throughout this project, I have gotten the opportunity to really grow in terms of being independent and responsible. Since the project did not have any definite requirements other than creating a dataset, I had a lot of creative freedom. I thought I always wanted this freedom, but I also realized how much responsibility it comes with and had to learn to deal with it. All of a sudden there were no defined criteria or obligations to fall upon so I had to learn to create these on my own in all stages of the project. I am very grateful to my advisor Dr. Guohao Lan for guiding me and repetitively reminding me to focus on the most important things and not get stuck on details or be overly critical of myself or set too ambitious goals given the duration of the project. I would also like to thank Lingyu Du for being so kind and helpful whenever I needed support in any practical matters. I am also incredibly thankful to Dr. Xucong Zhang for his valuable insights and for showing genuine care for my progress during the few discussions we had.

Bishwas Regmi
Delft, The Netherlands
19 April 2024

Contents

List of Figures

List of Tables

1 Introduction

1.1. Emotion Recognition

Emotions have a tremendous impact on our daily lives. Being able to recognize these emotions can be beneficial in many areas of life, from personal well-being to more natural and effective human-machine interactions. Emotions are often more genuine indicators of how a person feels, compared to what they might express verbally or consciously [1]. Recognizing these emotions can thus offer an authentic insight into an individual's state of mind, which could be invaluable in areas like healthcare, human resources, or even market research. Also, accurate emotion recognition can lead to more effective and empathetic communication in a clinical setting, for example, where understanding a patient's emotional state could provide clues to underlying issues that may not be immediately apparent.

Emotions naturally play a crucial role in our decision-making processes. By recognizing these emotional signals, one can make more informed choices, both at an individual level and at the level of larger organizations or systems. In the area of mental health, early identification of emotional states can lead to better mental health interventions. For instance, catching signs of stress or depression early on could lead to more effective treatments. As technology becomes increasingly integrated into our lives, systems that can recognize and adapt to human emotions could also offer more intuitive and satisfying user experiences. In high-stakes environments like driving or operating heavy machinery, emotion recognition can serve as a safety measure by generating alerts and notifications when an individual is in a state that could compromise their ability to control and operate safely.

Emotion recognition can be defined as the process of identifying human emotions through various means such as facial expressions, physiological signals, body language, voice patterns, etc. It is a multidisciplinary field stemming from the fields of Psychology, Computer Science and Neuroscience. The scientific study of emotions began as early as 1872 when Charles Darwin's book "The Expression of the Emotions in Man and Animals" was published, where it was proposed that emotional expressions were universal among humans. Building on Darwin's work, significant advancements have been made in understanding and classifying human emotions. Most notably, Paul Ekman developed the Facial Action Coding System (FACS) in the 1970s, which provided a detailed understanding of facial expressions and their links to emotions. As technology evolved in the areas of computer vision, speech analysis, and sensor technology, computational methods has enabled great advancements in the capabilities of emotion recognition systems. These technologies enabled the analysis of a broader range of emotional indicators beyond facial expressions, including voice tonality, body posture, and physiological responses such as heart rate and skin conductance [2] [3] [4] [5]. With the recent advancements in deep learning and artificial intelligence, there has been an increase in research focused on developing systems capable of interpreting complex emotional states from multi-modal data sources [6].

1.2. Virtual Reality (VR)

Virtual reality (VR) is a computer-generated simulation of an environment or situation that immerses the user in an artificial world. Using VR headsets, users can see, hear, and interact with the virtual 3D worlds created using VR systems. Given the immersive and interactive capabilities of VR, it can become

an interesting and useful platform for emotion recognition. By simulating environments and scenarios that evoke genuine emotional responses using VR, emotional cues can be observed and analyzed in realistic and controlled settings. This capability can be very useful in psychological research, therapeutic applications, and provision of natural user experiences, where understanding and responding to human emotions are essential [7]. In VR, users can be placed in various situations that would be hard to replicate in the real world, which allows researchers to predict how individuals might react in similar real-world situations without the risk of getting them into real danger or other lasting consequences.

The interactive nature of VR is also beneficial for user engagement, making participants more involved in the tasks or scenarios presented to them. So it could be used in education, where VR provides students with immersive learning experiences that adapt in real-time to their emotional responses. This would ensure that the content is engaging and actually fitting to their needs. And in user experience design, emotion recognition studies in VR environments could be used to gain valuable insights which could be used to develop better products and services that resonate with users on an emotional level. VR used in combination with emotion recognition can thus prove to be a powerful and useful tool to provide value to the betterment of human lives in many ways.

1.3. Multi-modal approach

Multi-modal approach in emotion recognition means combining various signals or indicators of emotions such as facial expressions, voice intonations, body language, and physiological responses to identify and interpret a person's emotional state. Using multi-modal signals to recognize emotions provides obvious benefits as it brings the systems closer to how humans interpret and understand emotional cues. McKeown et al. argue that such an approach enhances the system's effectiveness by mimicking the nuanced ways humans perceive and react to emotions [8].

Another advantage of the multi-modal approach is error mitigation. In situations where one modality may be compromised (for instance, facial recognition being less reliable in low lighting), other modalities like vocal cues or physiological measurements can fill the gap, ensuring consistent and accurate emotion recognition. Ros et al. demonstrated that incorporating multiple sensors could increase the reliability of emotion recognition, especially in complex environments like social robotics, where accurately reading human emotions is crucial for safety [9]. Thus it is apparent that taking the multi-modal approach is inevitable in order to make effective emotion recognition systems.

1.4. Current Situation and Problem Statement

In emotion recognition using VR, VR has so far been primarily used as the emotion elicitation tool while external, specialized measurements such as heart rate variability (HRV), electrodermal activity (EDA) and brain wave patterns through electroencephalography (EEG) are used to gather the data on emotions. Marín-Morales et al. highlighted in their article "Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing" that the techniques that are used to recognize emotions evoked using VR are mostly autonomic nervous system (ANS) measurements such as HRV(73.8%) and EDA(59.5%) [10] . Research on central nervous system (CNS) measurements using EEG started becoming more prominent since 2016. And since 2019, the research expanded to include non-physiological implicit measures, such as eye-tracking and behavioral patterns.

While ANS and CNS measurements have proven to be highly effective for emotion recognition, they also come with significant drawbacks. The first major drawback is the cost. The specialized equipment required to accurately measure HRV, EDA and EEG are often expensive, making them less accessible for widespread use in research and practical applications. The second drawback is the intrusiveness of these measurement equipment, as they require the attachment of sensors or electrodes to participants, which can lead to discomfort and altered or unnatural behavior. And the third major drawback is the need for a controlled environment to ensure the accuracy of these measurements, which adds complexity and limits the feasibility of conducting studies on emotion recognition outside of laboratory conditions.

The heavy reliance on external and specialized equipment fails to utilize VR's full capacity for a comprehensive approach to emotion recognition. VR is presently primarily used for emotion elicitation, while its inherent ability to track behavioral signals such as eye and head movements, which could be used

for emotion recognition is being overlooked. By fully leveraging VR's capabilities, not only as a tool to elicit emotions but also to recognize emotions using its capacity to measure behavioral signals, affordable and more accessible emotion recognition systems could be developed. These systems would also provide a comfortable and natural way of recognizing emotions, and could be more easily employed in real-world settings. Thus, the following statement is drawn to concisely summarize the identified current problems in the field of emotion recognition using VR.

Problem Statement: The use of VR in emotion recognition is limited to only eliciting emotions and there is a heavy reliance on external, costly, and intrusive measurement equipment. This reliance overlooks VR's inherent capacity to also function as a measurement tool, and offer non-intrusive and cost-effective solutions for emotion recognition.

1.5. Project Objectives

1.5.1. Proposed solution

The proposed solution to the identified problem is to shift from traditional ANS and CNS measurements like HRV, EDA, and EEG to using Virtual Reality (VR) together with behavioral measures such as eye and head movements for Emotion recognition. VR not only reduces the necessity for expensive external equipment because of its built-in eye and head-tracking features but also offers a much less intrusive experience for users. This makes capturing emotional responses not just easier and more affordable, but also more comfortable for participants. Also, VR's ability to integrate multi-modal sensory inputs—from tracking eye movements to head movements and recording voices—allows for a more contextual human-like interpretation of emotions.

The Facial Action Coding System (FACS), a comprehensive tool for objectively measuring facial movements for emotion recognition developed by Ekman and Friesen showed that subtle movements in the peri-ocular regions can be used to study emotional expressions [11]. It shows that a significant number of facial cues associated with emotion are concentrated around the eyes, with as much as 65% of the total facial Action Units (AUs) linked to emotional expression found in the region around the eyes. Further research has shown that eye expressions alone, such as pupil dilation and blink rate, offer a deep insight into a person's emotional state [12]. Likewise, research conducted by Samanta et al. highlights the significant role of head movements in conveying emotions and facilitating communication [13]. Furthermore, voice input provides valuable information on emotional states through variations in tone, pitch, and the pace of speech [14]. These studies suggest that the sensory data collection capabilities inherent to VR technology are sufficient for effective emotion recognition.

1.5.2. Main challenges

One of the primary challenges in the field of emotion recognition using VR is the lack of active and immersive stimuli that are specifically designed to fully utilize the capabilities of VR, and are validated to effectively evoke a wide range or emotions. Such stimuli to evoke even just the basic emotions identified by P. Ekman which are happiness, sadness, surprise, anger, disgust and fear are not publically available. Marín-Morales et al. (2020) observed that the majority of research (90.5%) on emotion recognition using VR has so far narrowly focused on arousal and on the emotions associated with high arousal, such as stress, anxiety, and fear [10]. While, valence—the positive or negative value attached to an emotional experience—and specific emotions associated with it, like happiness, sadness, and disgust, have not received much attention. Thus, there is a need for the creation of effective active stimuli that maximize the utilization of the immersive and engaging capacities of VR and are capable of eliciting a wide range of discrete emotions in order to fully explore and understand emotional responses within VR environments.

Another significant challenge is the scarcity of comprehensive emotion recognition datasets covering at least all the basic emotions. The field of emotion recognition using VR is in its first growth phase. The availability of such datasets is thus a limiting factor for researchers aiming to develop robust models, particularly deep learning algorithms, which depend on diverse and large volumes of data for training to ensure high accuracy and reliability. To push the research forward in this field, there is a need for the creation and publication of such comprehensive datasets available for all researchers.

1.5.3. Project scope and limitations

Addressing the main challenges identified previously, the first is the lack of publicly available, validated immersive (active) stimuli. This challenge largely falls outside of the author's direct expertise, given its roots in psychological and behavioral research. However, it is important to note that non-immersive (passive) stimuli, such as movie clips, while perhaps not as effective in fully delivering the dynamic and interactive nature of real emotional experiences as immersive stimuli do, have undergone extensive study and have been proven to be effective. These non-immersive stimuli, which are publicly available, present a viable alternative for the time being, offering a sufficient means to conduct emotion recognition research until more immersive and comprehensive stimuli become available.

The second challenge regarding the insufficiency of emotion recognition datasets available for widespread research more closely aligns with the area of the author's expertise. It is in this area where the author sees the opportunity to make the most significant and meaningful contribution by expanding and enriching the datasets available to the research community. Combining the proposed solution to the previously defined problem statement and the goal of tackling the identified second challenge, the following objective has been set for this project.

Objective: Use Virtual Reality (VR) to simultaneously elicit emotions and record the emotional responses to create a multi-modal dataset. The multi-modal dataset shall only contain signals that can be captured directly through a standard VR headset, which includes recording of the peri-ocular region, recording of the stimuli, tracking eye behavior (such as pupil responses and gaze patterns), recording of the audio, and recording of the head movements.

1.6. Related Work

There have been some works published in recent years that closely align with the objective of this project. Three of such works are summarized below and the areas in which this project can provide novel contributions are highlighted.

1.6.1. VREED Dataset (2021)

In December 2021, Tabbaa et. al. published the multi-modal affective dataset called VREED (VR Eyes: Emotions Dataset) [15]. The participants were shown movie clips using a VR headset while their behavioral and physiological data were being recorded. The recorded physiological data include heart activity (ECG) and skin response (GSR) while the behavioral data include gaze fixations, saccades, and blink rates.

Our Novel contribution

1. **Recording of the Peri-Ocular Regions**: This addition is particularly valuable for Facial Action Coding System (FACS) analysis. Also, deep learning techniques can potentially uncover subtle, previously unexplored features in the peri-ocular region, enhancing the current understanding of emotional expressions.
2. **Recording of Stimuli and Gaze-point**: Capturing the stimuli that evoke emotions, along with where participants are looking (gaze-point), offers critical insights into the emotional triggers, aiding in more accurately predicting the emotions being experienced.
3. **Inclusion of Pupil Diameter Data**: The diameter of the pupils provides a reliable indicator of arousal levels, offering a direct physiological measure of emotional intensity.
4. **Inclusion of Head Movement Data and Sound Recording**: Patterns in head movement and sound are closely tied to emotional expressions, providing additional dimensions to the emotional data collected.
5. **Exclusion of ECG and GSR Measurements**: The autonomic nervous system (ANS) measurements like ECG and GSR shall be excluded based on the considerations discussed earlier in section 1.4

1.6.2. EMO-Film Dataset (2022)

In April 2022, Zhao et al. introduced the EMO-Film Dataset[16]. The hardware used for creating the dataset was a custom smart-glasses frame equipped with one camera targeting the right eye and another capturing

the world view/stimuli, as illustrated in figure 1.1. The modalities included in the dataset are recordings of the right peri-ocular region, recording of the world + 2d gaze point, and right eye pupil diameter.

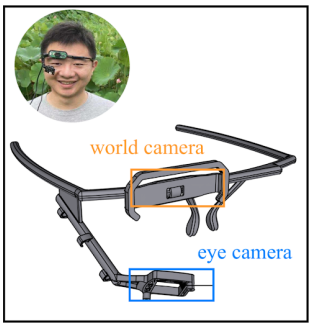


Figure 1.1: EMO-Film dataset collection hardware [16]

Our Novel contribution

1. **Transition to Virtual Reality (VR)**: VR offers a more controlled, consistent, and customizable environment for emotion recognition studies, enhancing the reliability and repeatability of the research.
2. **Recording of Both Eyes**: Capturing data from both eyes, instead of just one, provides a richer dataset.
3. **Inclusion of Head Movement Data and Sound Recording**: Patterns in head movement and sound are closely tied to emotional expressions, providing additional dimensions to the emotional data collected.
4. **3D Gaze Point Over 2D**: Upgrading from a 2D to a 3D gaze point offers a deeper level of information.
5. **Recording Eye Images at 120 fps**: Increasing the frame rate to 120 fps from 30 fps allows the dataset to capture the subtle microexpressions, which last only a fraction of a second but are highly informative regarding emotional states

1.6.3. CEAP-360VR Dataset (2021)

In December 2021, Xue et. al. released the CEAP-360VR dataset[17]. This dataset was created using a VR headset along with a wristband for measuring physiological signals and a joystick for continuous annotation of the perceived emotion. The modalities included in the dataset are pupillometry data, gaze data, head movement, heart rate(HR), skin temperature(SKT) and electrodermal activity (EDA).

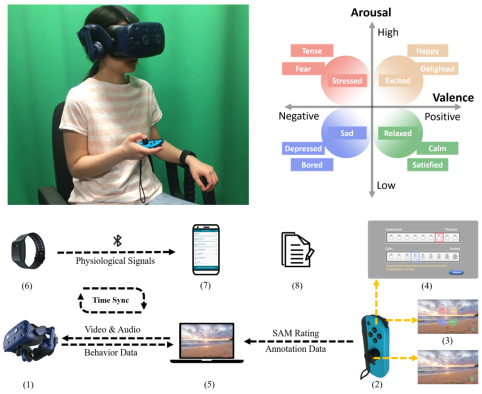


Figure 1.2: CEAP-360VR dataset collection setup [17]

Our novel contribution

1. **Recording of the Peri-Ocular Regions**: This addition is particularly valuable for Facial Action Coding System (FACS) analysis. Furthermore, deep learning techniques can potentially uncover

subtle, previously unexplored features in the peri-ocular region, enhancing the current understanding of emotional expressions.

2. **Recording of Stimuli and Gaze-point**: Capturing the stimuli that evoke emotions, along with where participants are looking (gaze-point), offers critical insights into the emotional triggers, aiding in more accurately predicting the emotions being experienced.
3. **Focus on Discrete Basic Emotions**: Shifting the emphasis to discrete basic emotions rather than Valence-Arousal levels allows for a more detailed exploration of emotional states.
4. **Elimination of Continuous Annotation Requirement**: Removing the need for participants to continuously annotate their emotions minimizes cognitive load, potentially leading to more natural and spontaneous emotional responses.
5. **Exclusion of Physiological Signal Recording**: Excluding the recording of physiological signals focuses the dataset on behavioral and visual cues, streamlining the analysis process and emphasizing cost-effective and more readily accessible modalities for emotion recognition.

1.7. Main Required Components

The creation of an emotion recognition dataset requires three main components: the stimuli, the data collection system and the labeling process.

The stimuli are the selected materials or experiences designed to elicit emotional responses from participants. The choice of stimuli plays an important role in creating a good dataset as it directly impacts the diversity and intensity of the emotions captured and thus influences the dataset's usability in further research. The data collection system encompasses the hardware and software used to record participants' emotional responses to the stimuli. The precision and usability of the data collection system determine the quality and reliability of the data gathered. The labeling process involves annotation of the collected raw data with labels that accurately describe the emotional states experienced by the participants. It is thus essential for transforming raw data into a structured dataset that can be used for training emotion recognition models. The selection and development of these components are discussed separately in the upcoming chapters.

1.8. Thesis Structure

In this chapter, an analysis of the current situation in the field of emotion recognition using VR has been done and a concise problem statement is formulated. Then a possible solution is proposed to the problem and the scope of the project is set, taking into consideration which other published works have also tried to solve the problem and what novel contributions can be made. The objective of the project is set to to create and validate a Multi-modal Emotion Recognition in VR Dataset. This involves selecting appropriate stimuli to reliably evoke specific emotions, designing an efficient data collection system, and establishing an effective labeling process. In chapter 2, the process of selecting the best available stimuli is discussed. In chapter 3, the design of the data collection system is described, detailing its components and functionalities. In chapter 4, different labeling processes are explored and their advantage and drawbacks are compared. In chapter 5, the created dataset is validated by evaluating the data collection process' effectiveness in eliciting intended emotions, as reported by participants, and by analyzing pupil diameter data to verify the presence of emotion-specific information within the dataset.

2

The Stimuli

The stimuli play an important role in the creation of a high-quality emotion recognition dataset because they directly impact the utility of the dataset. By evoking genuine emotional responses from the participants, the chosen stimuli enable the recording of information-rich data. If the stimuli are not effective at evoking the targeted emotions for study, the recorded data will lack the information necessary for meaningful research, making the dataset unreliable and useless. Thus it is essential to carefully select high-quality stimuli that have been empirically validated for their effectiveness in evoking a wide range of emotions.

2.1. Emotion models

The emotion models that are used to create labels for the emotion classes can be categorized into two main groups: Categorical and Dimensional [18]. The most widely used categorical model is Ekman's Six Basic Emotions and the most widely used dimensional model is Russell's Circumplex Model of Emotions. Ekman identified six basic innate emotions: Happiness, Sadness, Fear, Disgust, Anger and Surprise. Each with universal facial expressions that are recognized across different cultures. While Russell suggested that emotional states are interconnected rather than discrete and proposed a two-dimensional space in which emotions can be arranged within a circular space. The two primary dimensions are valence (pleasantness) and arousal. The relationship between the two models is shown in figure 2.1.

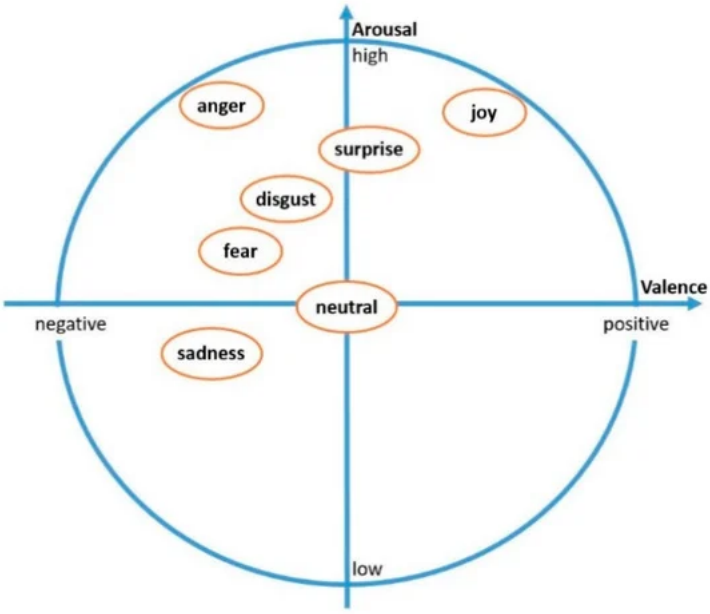


Figure 2.1: Ekman's Six Basic Emotions mapped on the 2D arousal-valence chart of Russels's Circumplex Model of Emotions [19]

The emotion model that has been chosen for the labeling of the dataset in this project is Ekman's Six Basic Emotions. This model has been chosen because it is based on the fact that the basic emotions have distinct identifiable facial expressions that can be universally recognized across different cultures. Since this model is associated with specific facial expressions, it is very well suited for a dataset focusing on capturing facial expressions. The discrete categorization of emotions simplifies the task of labeling the emotional responses in the dataset. Also, in practical daily applications, the ability to identify discrete emotional states would be more valuable than assessing levels of valence and arousal. However, acknowledging the subjectivity and variability of emotional experiences each of the six discrete emotions will be assigned a numerical rating representing the experienced emotional intensity by the participants during labeling.

2.2. Selection criteria

When selecting the stimuli database to be used in the data collection process, the following two criteria have been used:

1. The stimuli must be capable of triggering all of Ekman's six basic emotions.
2. The stimuli must be validated through scientific research.

The first criterion is used to comply with the selected emotion model, which is Ekman's six basic emotions. The reason for the selection of this emotion model is explained in section 2.1. The second criterion is used to ensure effectiveness in reliably evoking the intended emotional responses for the reasons provided at the beginning of the chapter. Scientific validation ensures that the selected stimuli are not only theoretically capable of inducing specific emotions but have also been scientifically proven to do so.

2.3. Stimuli types

In the context of emotion elicitation in virtual reality, stimuli presented to participants can be categorized into two types: Active and Passive. Active stimuli require participants to engage directly with the virtual environment, interacting with elements or completing tasks. Passive stimuli only require the participants to observe the presented content without any requirement for direct interaction.

2.3.1. Active

The active stimuli that are compatible with VR include in general VR games and custom Virtual Reality Environments (VREs) designed to evoke particular emotions. VR games would arguably be the most effective at evoking genuine emotional responses from participants as the participants would be fully immersed and engaged in the activity. However, because of the open and unpredictable nature of the games, it would be incredibly complex to keep track of the emotional triggers encountered by the participants and label the data containing the emotional responses effectively. So, having systematically designed stimuli where the researchers have knowledge of the points at which emotional responses are likely to be evoked, significantly simplifies the data collection process. This could be achieved through custom-designed VREs, where participants are guided through an immersive and engaging environment containing strategically placed triggers to evoke some intended emotions. This would provide a good balance between the immersive and engaging nature of VR games to evoke genuine emotions while maintaining the predictability and controllability to effectively collect and label useful data. Efforts are being made to develop such virtual environments to evoke some of the emotions [20] [21][22]. However, the creation of comprehensive virtual environments capable of evoking the full spectrum of basic emotions is still in the primary stage of research, thus no publicly available Virtual Environment stimuli dataset that meets the requirements was found at the time of this study.

2.3.2. Passive

The passive stimuli that can be adapted to VR are 360°panoramic videos, non-panoramic 2D videos, and images. While 360°-videos may not offer interactivity like Virtual environments, they can still provide an immersive experience. The only publicly available database of 360°videos for emotion elicitation that was

found during the time of this study was made available by Li et. al. [23]. The videos in the database however were only labeled with valence and arousal ratings and not with discrete emotions. Thus, this stimuli database did not satisfy the requirements of being validated for evoking all of the basic emotions. This meant that the non-panoramic 2D videos were the only options left. However, this doesn't imply a compromise in effectiveness. In fact, 2D video clips, which are usually extracted from movies, have been extensively studied by psychologists and proven to be effective at eliciting the targeted emotions since as early as 1993 [24]. The 2D video clips when viewed through the VR headset also seem to be more engaging and less distracting than the 360°videos. This is because the participants would always have the whole 2D video frame in their mid-peripheral field of vision. So they would not have to constantly look around in the virtual environment to grasp the full context of the presented scenario, which would be distracting.

2.4. Final selection

The stimuli that were chosen for this study were taken from a database consisting of movie clips assembled by Zupan et. al. [25]. Zupan et. al. addressed significant limitations inherent in previously validated stimuli databases that have been frequently utilized in emotion-related research. The limitations included outdatedness, a limited set of evoked emotions, and/or focus around only one emotion. The term "outdated" here implies that the audio-visual quality of these clips does not meet current expectations, and the depicted social contexts may no longer resonate with modern audiences because many clips included in the databases are over two decades old. Zupan et. al. thus aimed to ensure that the film clips used in research remain relevant by re-validating and extending the thoroughly studied film clips stimuli listed in table 2.1. Also, additional criteria were used for selecting the movie clips such as, the spoken language in the clips must be English, the duration must be less than 3 minutes, and the scenes must be easily understandable without much knowledge about the storyline and characters.

Table 2.1: List of publicly available emotion elicitation databases re-validated by Zupan et. al. [25] This list contains the stimuli databases that have been extensively used in emotion recognition research.

Study	# of Clips	# of Emotions	Emotions
Gross & Levenson (1995)	78	8	Amusement, Anger, Contentment, Disgust, Fear, Neutral, Sadness, Surprise
Schaeffer, Nils, Sanchez & Philippot (2010)	70	7	Amusement, Anger, Disgust, Fear, Neutral, Sadness, Tenderness
Gabert-Quillen, Bartolini, Abravanel & Sanislow (2015)	18	9	Amusement, Anger, Calmness, Disgust, Excitement, Fear, Happiness, Sadness, Surprise
Gilman, Shaheen, Nylocks, Halachoff, Chapman, Flynn…Coifman (2017)	15	10	Affection, Amusement, Anger, Disgust, Fear, Guilt, Happiness, Interest, Neutral Sadness, Surprise
Zupan & Babbage (2017)	30	5	Angry, Fearful, Happy, Neutral, Sad

For each of the basic emotions including neutral, 2 movie clips were chosen from the stimuli database. Zupan et. al. provided the elicitation index scores for each of the movie clips in the database as can be seen in Figures 2.2 till 2.15. These elicitation index scores were used to choose the most effective clips at evoking each of the basic emotions. It should be noted that since each movie clip can contain several scenes and dialogues, there can be some overlap between the emotions they evoke. For example, the clip which was selected for evoking disgust, was also rated quite high for evoking anger (see Figure 2.12). And, the clips which were selected for evoking anger, were also rated quite high for evoking disgust and sadness (see Figure 2.10 and 2.11). It is thus important to evaluate the emotional responses on the individual scene level within the video clips, rather than expecting the whole clip to evoke only one emotion.

2.5. Stimuli presentation

To preserve the effectiveness of the selected stimuli a few factors need to be considered. The stimuli clips were all extracted from movies. So participants had to be provided with the context of each clip to maximize their understanding and ensure the stimuli's effectiveness in eliciting the intended emotional responses. Without a proper understanding of the context, the impact of the stimuli might be compromised. To address this, participants were briefed with context sentences for each video clip before viewing. These context sentences were supplied by Zuphan et al. along with the dataset and were designed to prepare participants emotionally without diluting the intended effect. For the clips aimed at evoking surprise, the context sentences were slightly altered to maintain the element of surprise.

Also, the sequence in which the stimuli were presented required careful planning. Given that the stimuli were shown in quick succession, presenting a high-arousal emotion followed by a low-arousal clip could lead to unintended emotional carryover effects. Thus, a deliberate order was established: starting with Neutral (a), followed by Surprise (a), Neutral (b), Surprise (b), and then proceeding through Happiness (a,b), Sadness (a,b), Anger (a,b), Disgust (a,b), to Fear (a,b). The notation 'a' and 'b' imply clips with relatively lower and higher arousal levels, respectively. The first Surprise (a) clip was disguised as Neutral, and the Surprise (b) clip was disguised as Happiness, to preserve the unpredictability necessary for the surprise emotion.

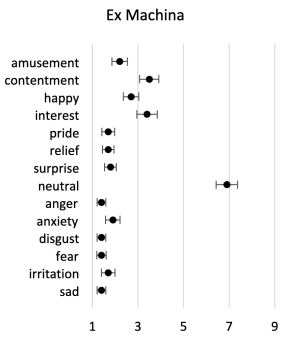


Figure 2.2: Elicitation ratings for clip: Neutral (a)

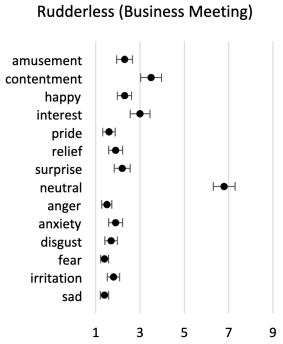


Figure 2.3: Elicitation ratings for clip: Neutral (b)

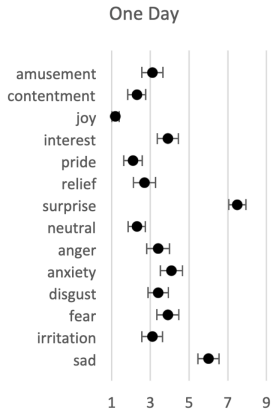


Figure 2.4: Elicitation ratings for clip: Surprise (a)

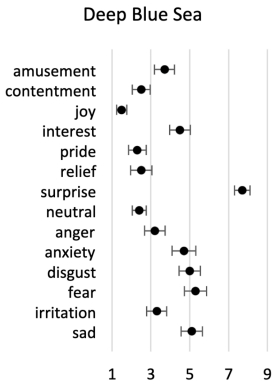


Figure 2.5: Elicitation ratings for clip: Surprise (b)

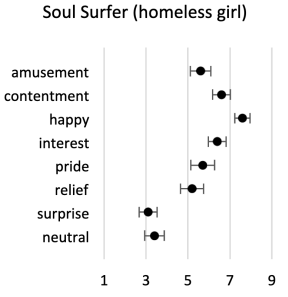


Figure 2.6: Elicitation ratings for clip: Happiness (a)

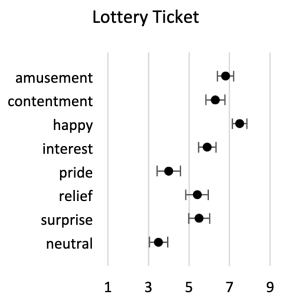


Figure 2.7: Elicitation ratings for clip: Happiness (b)

Figure 2.8: Elicitation ratings for clip: Sadness (a)

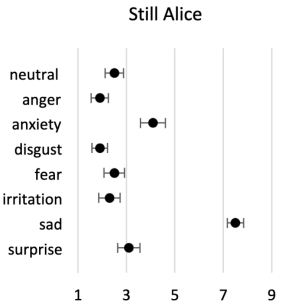


Figure 2.9: Elicitation ratings for clip: Sadness (b)

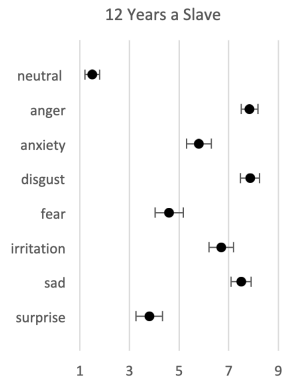


Figure 2.10: Elicitation ratings for clip: Anger (a)

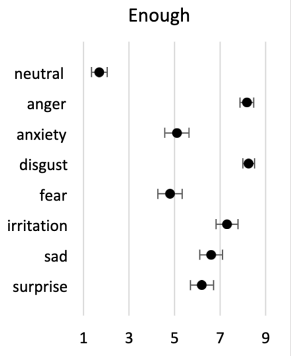


Figure 2.11: Elicitation ratings for clip: Anger (b)

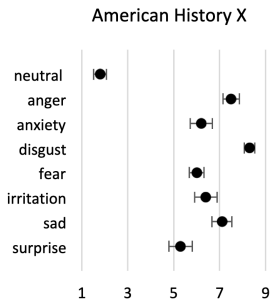


Figure 2.12: Elicitation ratings for clip: Disgust (a)

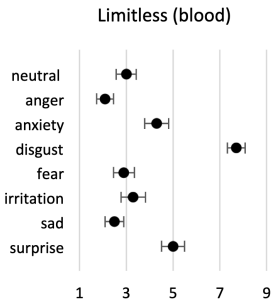


Figure 2.13: Elicitation ratings for clip: Disgust (b)

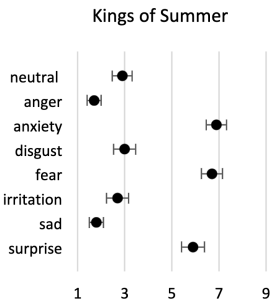


Figure 2.14: Elicitation ratings for clip: Fear (a)

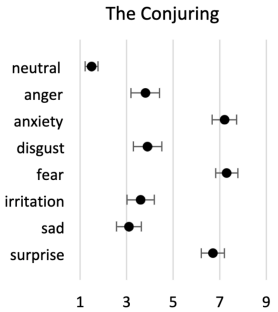


Figure 2.15: Elicitation ratings for clip: Fear (b)

3

The Data Collection System

The data collection system encompasses all the hardware and software components used for eliciting emotional responses from participants, recording these responses, and then labeling and processing the data to compile the final dataset. In this chapter, the components and their functionalities and interconnections are detailed. When designing the different components, a systematic approach is followed to ensure their proper functionality and efficiency. First, the user scenarios are outlined to understand the context in which the components need to operate. Based on the scenarios, specific requirements are identified and listed. Then, examining the listed requirements, the key design considerations are made where needed. And finally, the components are implemented to satisfy the identified requirements and design considerations.

3.1. Data collection Hardware

The primary choice for the data collection hardware was the 'VIVE Pro EYE' VR headset. This headset has integrated eye-tracking and recording capabilities along with other sensors such as IMU and microphones, satisfying the need to capture all of the intended data modalities through a single device. It is also one of the most popular and high-performance commercially available devices in its category and thus provides the best available resolution and sensitivity in capturing data modalities.

Figure 3.1: VIVE Pro VR headset

Figure 3.2: Pupil Labs binocular eye tracking add-on for VIVE Pro

However, accessing the raw eye images from the VIVE Pro EYE's integrated eye-tracking module was not possible because it required a special license for access. This is due to privacy concerns as the eye images contain iris data, which is classified as personally identifiable information. Therefore the 'VIVE Pro' VR headset, a model similar to the VIVE Pro EYE but without the integrated eye-tracking, was selected as

an alternative (Figure 3.1). This was then paired with an external eye-tracking add-on from Pupil Labs (Figure 3.2) to enable the recording of raw eye images without the need for special licensing.

3.2. Software Infrastructure

3.2.1. Tools and frameworks

Below is an overview of all the software tools and frameworks used in the development of the software components for this project. For each tool or framework, their specific function and the reason for their selection are provided.

Virtual Environment Development

Unity360 Unity360 is an extension of the Unity engine which specializes in the creation and deployment of immersive Virtual Reality Experiences. Unity360 was chosen as the foundational platform for the project because it is a leading platform for VR development and thus has wide community support, offers compatibility with a broad range of VR hardware, and offers a wide array of tools and features for the development of VR applications. Another key consideration for this choice was the exclusive availability of Pupil Labs' 'hmd-eyes' package within Unity. This package provides resources that can be used to communicate with the Pupil Labs' software and integrate the eye-tacking capabilities directly into the VR environment.

OpenXR Plugin OpenXR plugin for Unity is a software package that implements the OpenXR API within the development environment. OpenXR is a unified open-source API standard aiming to simplify VR and AR development by providing a common API that can be used across a wide range of hardware platforms. This enables one to create applications compatible with any VR device that adheres to the OpenXR standard, significantly reducing the need for device-specific programming. It has been chosen to use this standard mainly for the future-proofing of the application. By using OpenXR, the application remains flexible and can be easily adapted to newer, more advanced VR hardware if the need arises to replace the current hardware.

Programming Languages The primary programming language used for the development of the virtual environment and the recording process is C#, given Unity's exclusive support for it. In addition, Python has been used to write automation scripts that manage tasks such as invoking external tools and services, as well as processing and feeding data into them. The Python scripts are executed from within the Unity environment, using the System.Diagnostics.Process class in C#.

External Tools and Services

Pupil Capture Pupil Capture is an open-source software developed by Pupil Labs. It is designed to enable comprehensive eye-tracking data collection and real-time analysis. Pupil Capture interfaces with the eye-tracking hardware to capture high-resolution eye images. It then extracts from them detailed eye-tracking metrics such as pupil size, gaze positions, blink rate, fixations, etc.

Through the Network API, Pupil Capture connects with Unity's hmd-eyes package. The Network API offers two main communication channels: 'Pupil Remote' for text-based commands and 'IPC Backbone' for real-time data streaming. Pupil Remote enables remote control over recordings and calibrations from within a Unity application via ZeroMQ's REQ-REP. The IPC Backbone enables real-time access to nearly all data generated by Pupil Capture including the eye-images stream from within Unity application via ZeroMQ's PUB-SUB pattern. This setup enables monitoring of the recorded eye images within the Unity application and transmission of the scene video and head movement data from the VR environment to Pupil Capture, ensuring synchronous data collection and storage.

Pupil Player Pupil Player is a post-processing software by Pupil Labs designed for analyzing and visualizing eye-tracking data recorded with Pupil Capture. It allows users to review recordings, overlay gaze data on the scene recordings for visualization, and more importantly, process raw data into refined

metrics. These metrics can then be exported in more general formats so that they can be further analyzed with external tools.

Label Studio Label Studio is an open-source data labeling tool designed for the annotation of diverse types of data, such as images, audio files, texts, and videos. It supports a wide array of labeling tasks including classification, tagging, bounding box identification, etc. It has been chosen for this project because of its open-source nature, flexibility, and capacity to integrate with various data processing workflows and external tools. It offers customizable project templates which allow to design a tailored labeling interface according to specific requirements. Furthermore, it provides a comprehensive API which enables a seamless synchronization between the data generated within the Unity application and the Label Studio platform. This direct API integration streamlines and simplifies the workflow between data capture, processing and labeling.

Data Processing and Analysis

Jupyter Notebooks and Python For the project's data processing and analysis, Jupyter Notebooks and Python have been used for their interactive computing environment, simplicity, extensive libraries, and support for numerous data science and machine learning frameworks. For Deep Learning tasks, Pytorch framework has been used because of its flexibility and popularity in the research community.

3.2.2. Software architecture

The architecture of the data collection system is illustrated in Figure 3.3 using the 3-Tier Software Architecture model. The Presentation Tier contains the components with which the users directly interact

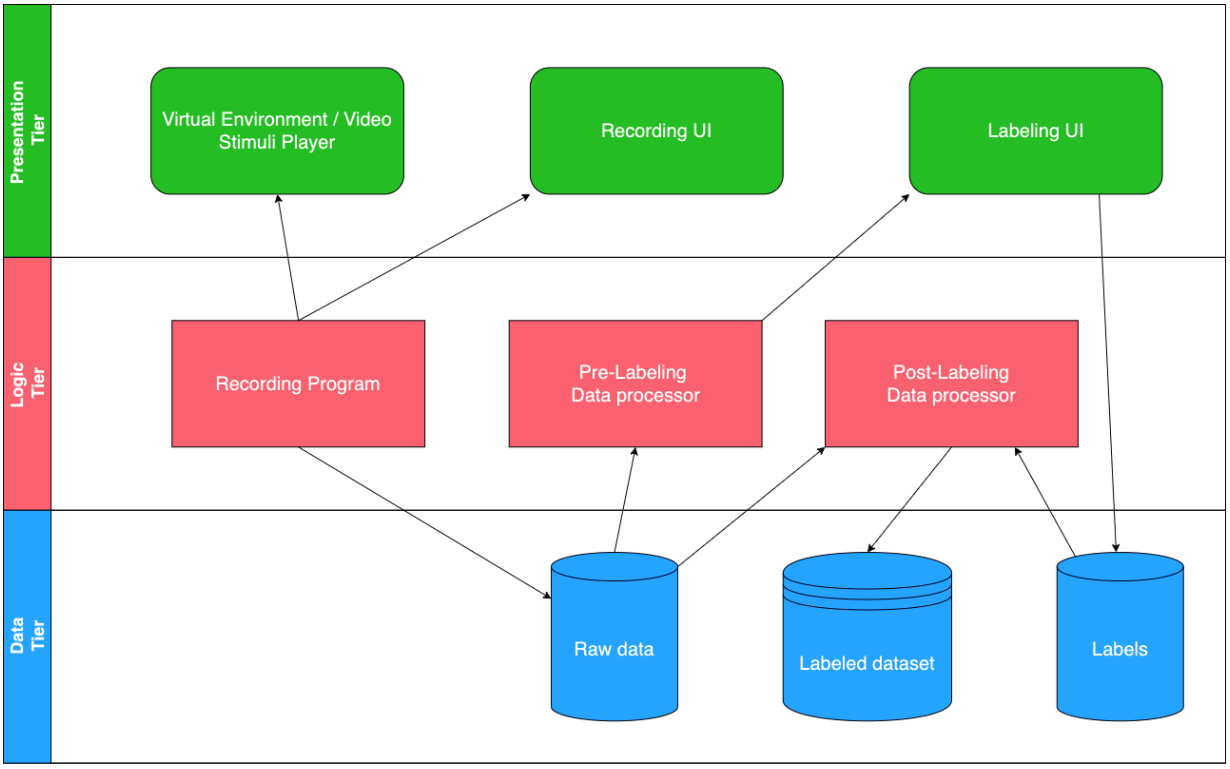


Figure 3.3: 3-Tier Software Architecture of the Data Collection System. The Virtual Environment is where the participant is presented with the stimuli videos. The Recording UI is used by the researcher to create and manage recording sessions. The Labeling UI is used to review the stimuli videos and annotate the segments. The Recording Program operates as the backend of the RecordingUI and the stimuli video player. The Pre-Labeling Data processor converts the raw data into a suitable format and presents it in the LabelingUI. The Post-Labeling Data processor combines the raw data and the labels to create the final labeled dataset.

with such as the Virtual Environment/Video Stimuli Player, Recording UI and the Labeling UI. The Logic Tier contains the core functional logic components that coordinate the data flow and processing in the system and includes the Recording Program, Pre- and Post-Labeling Data processing modules. And, the Data Tier contains components responsible for data storage and includes recorded Raw data, annotated Labels, and the final Labeled dataset. The function and implementation of these components are described individually in sections 3.2.4 through 3.2.9.

3.2.3. Data Collection Pipeline

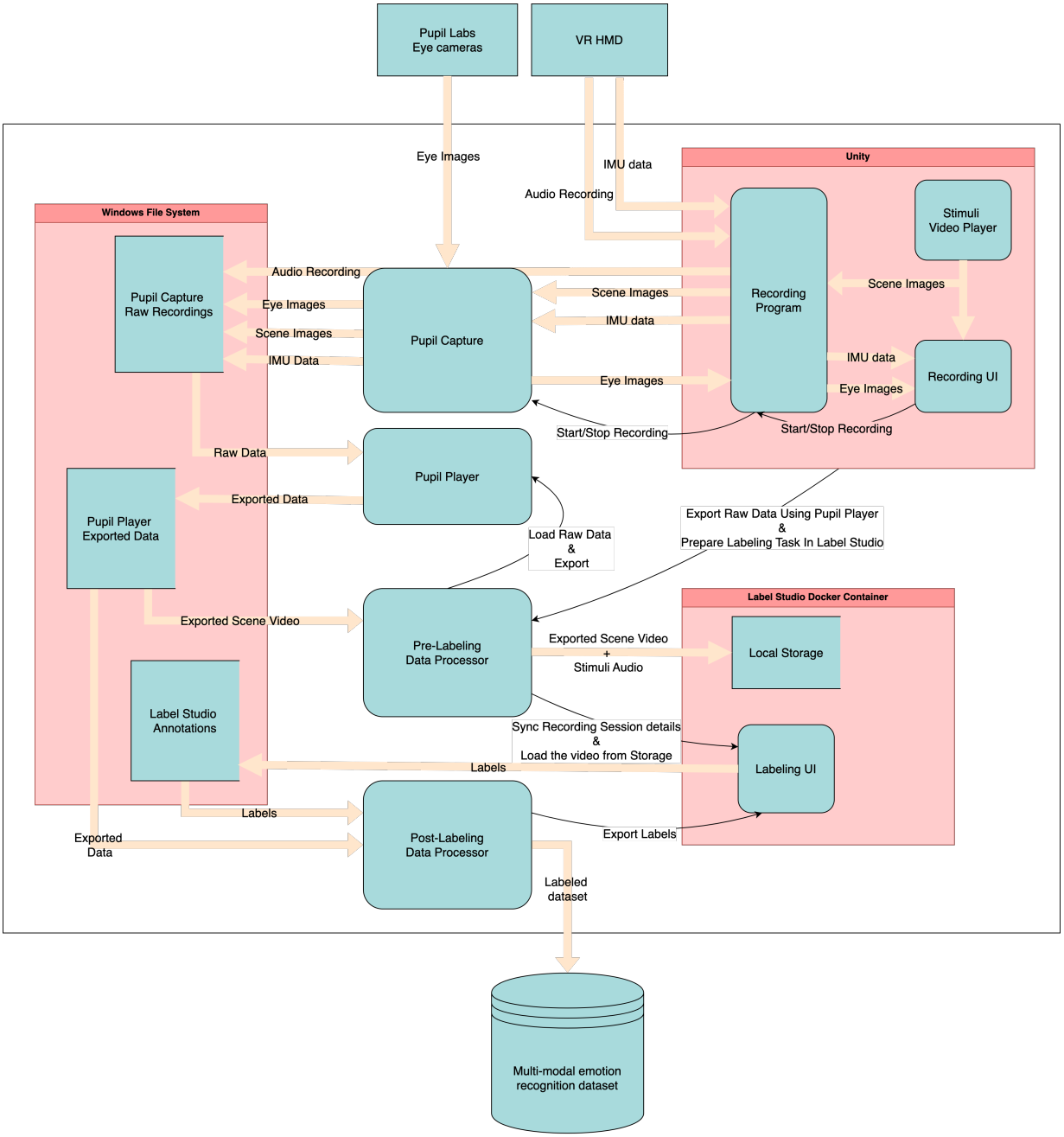


Figure 3.4: Full data collection pipeline. This diagram illustrates how the different data modalities are captured and handled by the different components in the data collection system to compile the final dataset.

Figure 3.4 illustrates the full data pipeline showing how the data generated from the Eye cameras and the VR headset goes through the different software components to eventually end up in the Multi-modal emotion recognition dataset. In the upcoming sections, the design and implementation of the software components that control this flow is described individually.

3.2.4. Virtual Environment

The virtual environment is where the participants find themselves when wearing the VR headset. The participants find themselves in this environment throughout the data collection process.

Usage Scenario

1. The participant wears the VR headset and finds themselves in the virtual environment.
2. The participant locates the stimuli video and positions themselves to view it.
3. When the video has finished playing, the labeling UI appears in the virtual environment.
4. The participant and the researcher annotate the collected data together using the labeling UI.

Requirements

1. Must be capable of playing the 2D stimuli videos.
2. Must have a labeling UI that can be viewed by both participant and researcher simultaneously.

Design Considerations

In order to provide a comfortable experience for the participants and make the data collection process efficient, a few considerations have been made during the design of the Virtual Environment. The first key consideration is the viewing experience of the stimuli videos. This should be as natural as possible to fully immerse participants and elicit genuine emotional responses. The next key consideration is the ease with which participants can comprehend the context of the videos, as this is essential for evoking the intended emotional reactions effectively. The presentation of the video stimuli should focus on maximizing the ease with which the participant get a clear understanding of the content.

Implementation

To address the first requirement of playing the 2D stimuli videos within the virtual environment, a video player with curved display has been implemented. A curved display was chosen because it mimics the

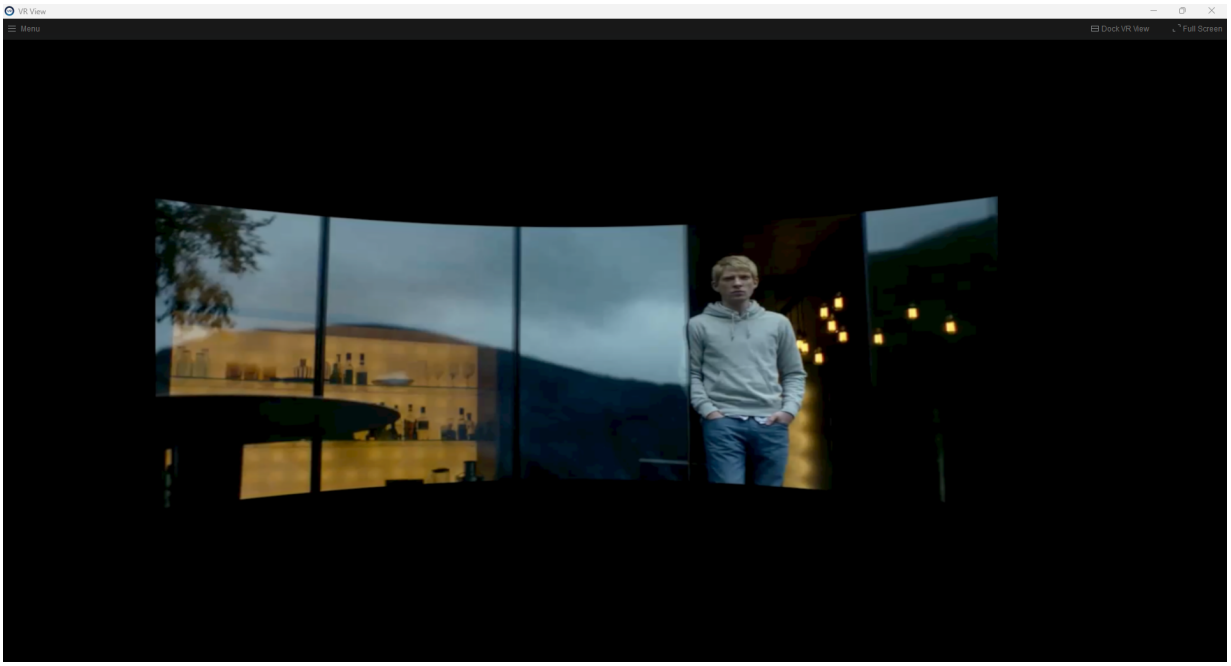


Figure 3.5: Participant's view of a video being played in the Virtual Environment.

natural curvature of the human sight providing a more natural and immersive viewing experience [26]. A snapshot of the participant's view of the video is shown in Figure 3.5.

To present the video in a way that maximizes the ease of understanding the content and context of the video, the dimensions of the video player display has been adjusted to fit within the mid-peripheral field of vision of the participants. The mid-peripheral field of vision for humans is approximately 120°horizontally and 60°vertically [27]. By doing so, the viewer is aware of what is happening in the entire video frame and can quickly shift focus within the frame by only moving the eyes and not having to move the entire head constantly. It was found during the initial testing of the video player in the virtual environment that having the video frame exceed the mid-peripheral field of vision resulted in a constant movement of the head which was distracting and was detrimental to the understanding of the video's context. This problem was fixed by adjusting the display size to fit within the approximate mid-peripheral field of vision.

To address the second requirement, a 'Display Duplication' feature has been integrated within the virtual environment, allowing the participants to view a duplicate of the computer screen used by the researcher. With this simple implementation, the researcher and participant can collaborately perform the annotation tasks without having to take off the VR headset after each video. A snapshot of the 'Display Duplication' feature is shown in Figure 3.6.

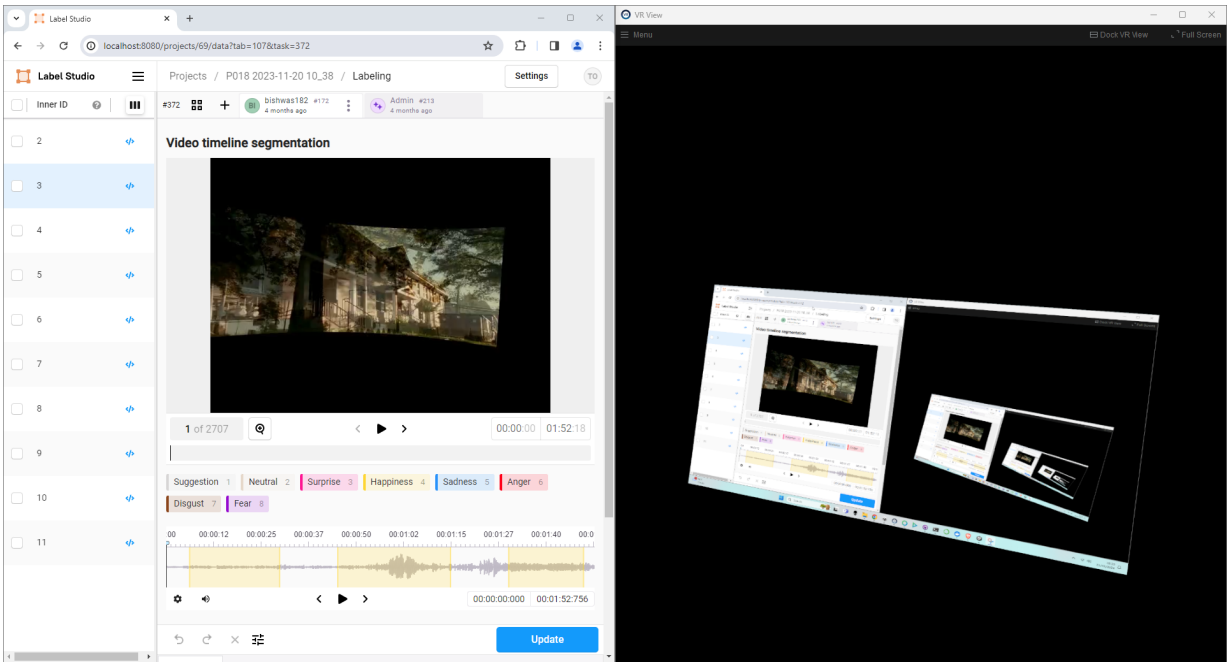


Figure 3.6: Researcher's view of the Labeling UI on a computer screen (left). And Participant's view of the LabelingUI using the Display Duplication feature implemented in the virtual environment (right). Note that the infinite mirroring effect is just an artifact of taking a screenshot. The participant sees the researcher's screen only once in reality.

3.2.5. Recording UI

The Recording UI allows for the control and monitoring of the data recording processes. It is used by the researcher to create and manage recording sessions, play the stimuli videos, initiate recordings and oversee all the signals being recorded while the participant is engaged in the Virtual Environment.

Usage Scenario 1

1. The researcher enters the participant's ID to create a recording session.
2. The researcher starts the eye-tracker calibration process.
3. The researcher selects a stimuli video to play.
4. The researcher confirms that all the recording components are connected and ready to record.

5. The researcher starts the video player and the data recording simultaneously.
6. The researcher monitors the connection status of the connected components and also the streaming of the recorded data.
7. After the video finishes playing, the researcher prepares the recorded data for labeling.

Usage Scenario 2

1. During the recording session, the participant needs to take off the VR headset.
2. The researcher stops the video and data recording.
3. The participant takes off the headset and put it back again.
4. The researcher re-calibrates the eye-tracker and continues the session.

Requirements

1. Must allow the creation and management of the session directory
2. Must enable selection and playback or stopping of a stimuli video from a pre-selected list.
3. Must record data concurrently during video playback and stop and save data when playback stops.
4. Must offer an option for re-calibrating the eye-tracking software without restarting the session.
5. Must enable pre-processing of the recorded data and preparation of the Labeling UI.
6. Must allow visual monitoring of all data streams being recorded during the data collection process.
7. Must enable monitoring of the recording program's internal state and the status of all connected hardware and software tools necessary for recording.

Design Considerations

A key consideration in designing the Recording UI is that all the monitoring signals should be clearly visible to the researcher at all times and if something is not functioning, it should be easily noticable. This could be done by color coding the monitored signals.

Implementation

Taking all of the requirements and design considerations into account, the final implemented design of the Recording UI, including an overview of its elements is shown in Figure 3.7.

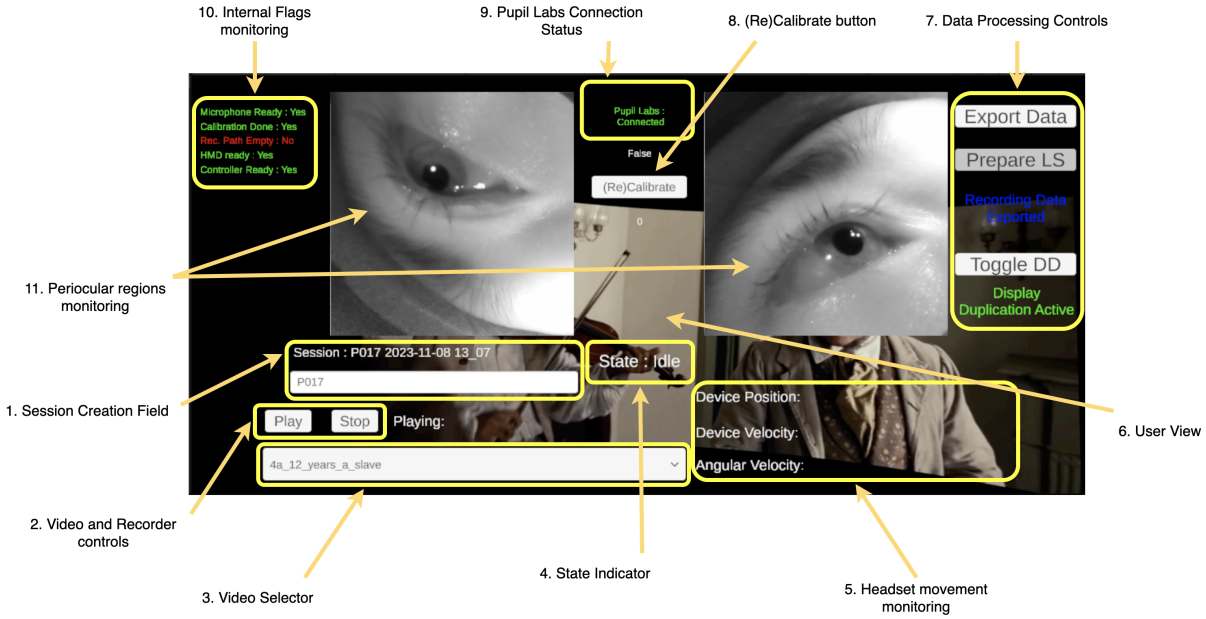


Figure 3.7: The Recording UI and its 11 elements. The researcher manages and monitors the recording sessions using this UI. This UI is only visible to the researcher.

To meet the first requirement, a session creation field (element 1) has been integrated where the participant ID can be entered. When the participant ID is submitted, a session directory is automatically created with the ID and the date and time of the creation. When a stimuli video is selected from element 3, a sub-directory named after the video is created within the session directory and the recording path is set to this sub-directory. The 'Rec. Path Empty' flag in element 10 indicates whether the sub-directory already contains recorded data when a video is selected. If the same video is played while the video sub-directory already contains data, an option is presented to either overwrite the existing data or cancel the operation.

To meet the second and third requirements, elements 2 and 3 have been implemented. A stimuli video can be selected from the dropdown list in element 3 and can be played using the 'Play' button in element 2. Upon clicking the 'Play' button, the stimuli video is played in the virtual environment and at the same time the recording program is also started in the background. When the 'Stop' button is clicked, the video playback and recording are stopped and the recorded data is stored. The 'Play' button and the dropdown list can only be clicked when the Recording program is in IDLE state.

To meet the fourth requirement, a 'Re-calibrate' button has been integrated into the UI as indicated by element 8. This button can also only be pressed when the Recording Program is in IDLE state.

To meet the fifth requirement, the 'Export Data' button and the 'Prepare LS' button in element 7 have been integrated. The 'Export Data' button automatically loads the raw data into Pupil Player in the background and exports the data into a suitable format for the labeling UI. The 'Prepare LS' button synchronizes the session and video names into its platform and loads the exported data and creates an annotation task.

To meet the sixth requirement, elements 5, 6, and 11 have been integrated into the Recording UI. In element 5, the IMU data of the VR headset is displayed. In element 6, the scene image as viewed by the participant is displayed. And, in element 11, the images of both eyes are displayed. An indicator of the microphone activity has also been implemented as a sound wave which only appears under element 8 whenever sound is detected.

To meet the seventh requirement, elements 4, 9 and 10 have been integrated. Element 4 indicates the State of the Recording Program. Element 9 indicated the connection status of the Pupil Capture software which provides the stream of eye images. And, element 10 indicates the connection status of the microphone and the VR headset(HMD), along with indicators for whether calibration has already been done and whether the recording directory of the currently selected video is empty.

3.2.6. Recording Program

The Recording Program coordinates the data capture from the eye cameras and VR headset, and the exchange of commands and information between the components the Presentation Tier and Data Tier such as the Recording UI, Virtual Environment, external recording tools and the data storage. Its processes are controlled from the Recording UI as it operates as the backend of the Recording UI. Its main tasks are to respond to the session management actions performed in the Recording UI and to coordinate the data flow during the recording process to ensure a synchronized collection of the data from different data streams. The session management tasks that it must respond to have been covered in the previous section of the Recording UI component. The data flow that it must manage during the recording process is illustrated in Figure 3.8.

Usage Scenario

1. The Recording Program receives the participant's ID from Recording UI and creates a new recording session.
2. The Recording Program receives the "Start Recording" command and initiates recording of audio, IMU data and scene images. Since the eye images are recorded externally through Pupil Capture, it notifies Pupil Capture to start recording as well.
3. The Recording Program displays the recorded data on the Recording UI. It receives eye images from Pupil Capture and also displays them in the Recording UI.
4. The Recording Program receives the "Stop Recording" command when the video is finished playing or stopped and forwards the command to Pupil Player and stops recording all data.

5. The Recording Program ensures that all the recorded data is stored at one place.

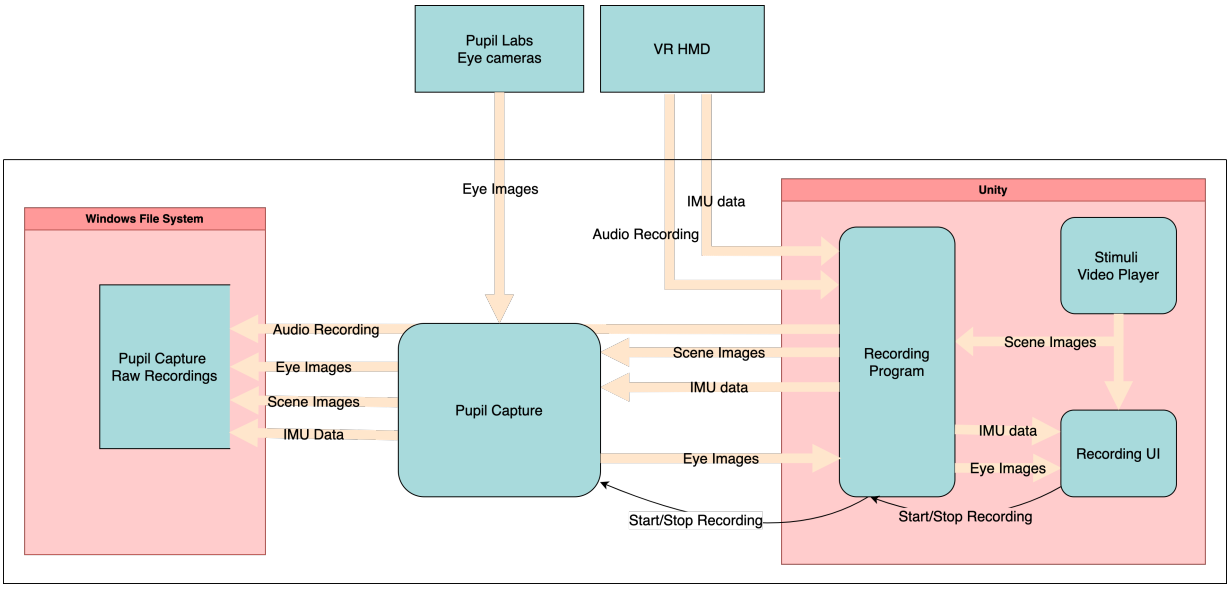


Figure 3.8: Data Flow managed by the Recording Program. The Recording Program is responsible for coordinating the recording of raw data from different modalities and storing them in one place.

Requirements

1. Must respond to session management actions initiated through the Recording UI.
2. Upon receiving the “Start Recording” command, must relay this command to Pupil Capture to initiate eye image retrieval and ensure these images are displayed on the Recording UI for monitoring.
3. Must simultaneously retrieve and display scene images from the Stimuli Video Player, and IMU data and audio recordings from the VR headset (HMD), on the Recording UI.
4. Must ensure synchronized recording and storage of Scene Images, IMU data, audio recordings, and Eye Images collectively in a single location.

Design Considerations

The Recording Program is at the core of the data collection process so it is essential that it performs reliably. To ensure its reliable operation, it can be structured as a Finite State Machine. This would restrict the program to a set of well-defined states and reduce any chances of it getting stuck in an indeterminate state. The next important thing to consider is the efficient use of the communication channel between Pupil Capture and the Recording Program in Unity. Since the eye images are being recorded directly by Pupil Capture and are being sent to Unity for monitoring purposes only, the transmission rate can be set at a much lower rate of 20Hz instead of the original recording rate of 120Hz to reduce unnecessary use of the communication channel. Also, the scene images do not need to be recorded at the same high rate as the eye images because the original recording rate of the stimuli videos ranges between 30Hz and 60Hz, so a higher recording rate would not be able to capture more information. It is essential to record the eye images at a high rate in order to capture the subtle involuntary eye movements, which is not the case for the scene images.

Implementation

In order to meet the first requirement of handling the session management actions performed in the Recording UI, a Finite State Machine as shown in Figure 3.9 has been implemented. The ’Idle’ state waits for user input or the creation of a new session. If a new session name is submitted, ’Creating New Session Directory’ state initiates the formation of a new session and its directory structure. ’Preparing Recording Path’ state prepares for recording by creating subdirectories based on the chosen video name and updating the ’Recording Path’ variable. If the ’Play’ button is clicked after session creation, ’Checking Calibration Done’ state verifies the completion of eye-tracking calibration, leading to the ’Prompting Calibration Option’

state, which presents the user with the option to perform the calibration or cancel. The 'Calibrating' state performs the eye-tracking calibration choreography and sends the calibration data to Pupil Capture once completed successfully. The 'Checking Recording Path Empty' checks whether the recording directory pointed by the 'Recording Path' variable is empty. If the directory is not empty, the 'Prompting Replace Recording Files' state asks the user whether to replace existing files in the recording path. If chosen to replace the existing files, the 'Deleting Recording Files' state proceeds to remove files. The 'Recording' state runs all the processes required for capturing the eye images, scene images, IMU data and microphone data. After the video has finished playing or is manually stopped, the 'Stop Recording and Store Data' terminates the recording processes and stores the collected data.

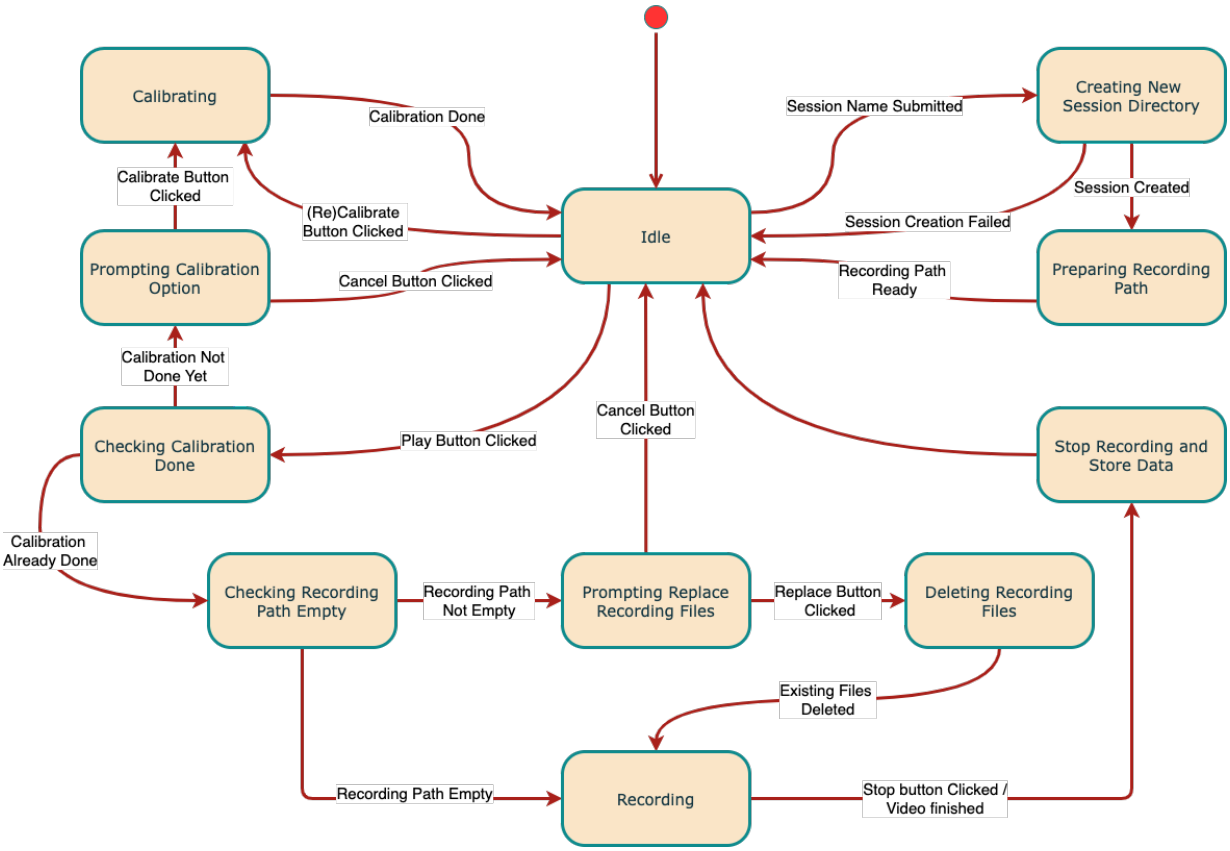


Figure 3.9: FSM diagram of the Recording Program showing its possible states and the transitions between the states.

To meet the second requirement, communication between the Recording Program in Unity and Pupil Capture had to be established. This has been done using the Network API which provides two communication channels: 'Pupil Remote' and 'IPC Backbone' using the ZeroMQ messaging library. The Pupil Remote channel utilizes the REQ-REP messaging pattern and is used to send the 'Start Recording' command, as well as the Recording Path generated in the Recording UI. The Recording Program then receives the eye images through the IPC Backbone channel utilizing the PUB-SUB pattern. Eye images are sent to Unity at a reduced rate of 20Hz, as opposed to the actual recording rate of 120Hz, because they are intended for monitoring purposes only, thus reducing unnecessary use of the communication channel.

To meet the third requirement of simultaneously retrieving the scene images, IMU data and audio recordings, multiple MonoBehavior GameObjects have been created in Unity that each handle one task. The ScreenCast GameObject captures the frames that are being shown through the VR headset and displays them on the Recording UI. The HeadTracker GameObject retrieves the IMU data from the VR headset which features the device's velocity, rotation and position. And, the MicRecorder GameObject records the audio detected by the integrated microphones in the VR headset.

To meet the fourth requirement of synchronized data recording and storing all the data collectively in a

single location, Pupil Capture has been chosen as the central recording tool. Pupil Capture already has the recording and storing feature available and can be easily controlled by sending commands through the Network API. The data gathered by the Recording Program in Unity is transmitted to Pupil Capture via the IPC Backbone channel of the Network API. The benefit of storing all the data through Pupil Capture is that it generates and stores the timestamps of each data entry sent to it, which is crucial for synchronizing data from multiple sources. While the eye images are set to be recorded at a rate of 120 Hz, the eye-tracking cameras function as independent, non-synchronized video sources, which means that a consistent sampling rate is not guaranteed [28]. The use of timestamps thus becomes essential for aligning the left and right eye images between themselves and with the other different types of data during analysis.

3.2.7. LabelingUI

After the participant has watched a video, the Labeling UI is used to review the video, and create and annotate segments within the video. The annotation is done by the participant and researcher together.

Usage Scenario

1. The researcher seeks through the stimuli video which was just played in the virtual environment.
2. The participant identifies the scenes where they clearly felt an emotion.
3. The researcher creates a segment in the video timeline and asks the participant to define its starting and ending points.
4. The participant gives a numerical rating indicating the emotion intensity for a segment and the researcher annotates that segment.

Requirements

1. Must be able to replay stimuli videos
2. Must allow easy creation and resizing of segments in the video timeline
3. Must allow individual labeling of the created timeline segments
4. Must allow assignment of numerical ratings to the segments

Design Considerations

After recording, the data must be properly annotated to ensure that the ground truth is as accurate as possible. Applying a single emotion class to all of the data recorded during a video session would be inaccurate, as a range of stimuli over the span of the video may elicit different emotions at different intensities. Thus, it must be possible to annotate individual scenes or segments within the video to enhance the granularity of the annotations. Doing so would enrich the dataset's quality by providing more precise ground truth labels.

Furthermore, loading of every data modality into the LabelingUI is not necessary because they are synchronized by their timestamps. By loading only the stimuli videos, the timestamps corresponding to all data modalities can be annotated. Thus the LabelingUI must be able to replay the stimuli videos and allow segmentation and labeling of its timeline. It must also allow the segments to be labeled as any of the available emotion classes and enable the assignment of a numerical rating to indicate the intensity of each emotional response.

Implementation

Instead of making a custom Labeling UI from scratch, Label Studio has been chosen and modified to fit the requirements. Label Studio is an open-source labeling tool that supports video annotation, provides customizable project templates for designing a tailored labeling interface, and has a comprehensive API that can be used to streamline and simplify the workflow between the data recording and labeling.

The Labeling Interface implemented using the Video Timeline Segmentation template within Label Studio is illustrated in Figure 3.10. It consists of a video player marked as element 1 in the figure, which satisfies the first requirement. The interface meets the second requirement through the functionality of element 2, which allows easy creation and resizing of segments on the video timeline. Segments can be created by simply clicking and dragging on the timeline. They can be resized by selecting the segments and dragging on either side of the segment box. To satisfy the third requirement, each segment can be

assigned an emotion class by selecting the segment and choosing from the labels shown in element 3. After assigning the emotion class for a segment, a numerical rating can be entered in element 4 to indicate the intensity of the emotion experienced during that emotion, satisfying the final requirement.

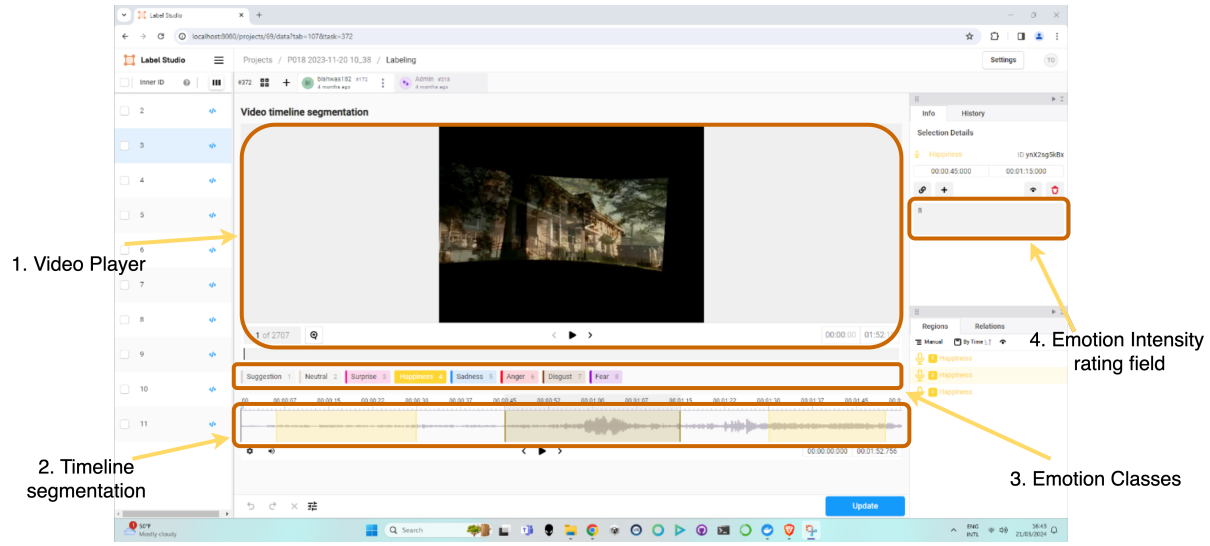


Figure 3.10: The Labeling UI and its main elements. Using the labeling UI, the participants can replay the stimuli video and select or create segments in the video timeline and annotate them.

3.2.8. Pre-Labeling Data Processor

Before labeling, the raw data collected by the recording program needs some initial processing to be accepted by the Labeling UI. The Pre-Labeling Data Processor converts the raw data into a format suitable for the Labeling UI and prepares the UI for annotation tasks. The data flow that the Pre-Labeling Data Processor must manage is illustrated in Figure 3.11 to provide a clear overview of the inputs and outputs for each component with which it interacts.

Usage Scenario

1. After a stimuli video has finished playing and the recording data is saved, the Researcher clicks on the ’Export Data’ button.
2. The Pre-Labeling Data Processor opens the external tool Pupil Player and loads the recorded data into it.
3. The Pre-Labeling Data Processor performs the ’Export’ operation within Pupil Player which converts the raw recorded data format into more general-purpose formats such as .mp4 and .csv files.
4. The researcher presses the ’Prepare LS’ button to proceed with data preparation for labeling.
5. The Pre-Labeling Data Processor transfers the exported scene video into the Label Studio’s local storage.
6. The Pre-Labeling Data Processor creates a project in Label Studio with the same name as the session name in Unity, if it doesn’t already exist.
7. The Pre-Labeling Data Processor then adds a new annotation task in the project and loads the exported scene video from its local storage into the labeling interface.
8. The Researcher and participant proceed to annotate the video timeline.

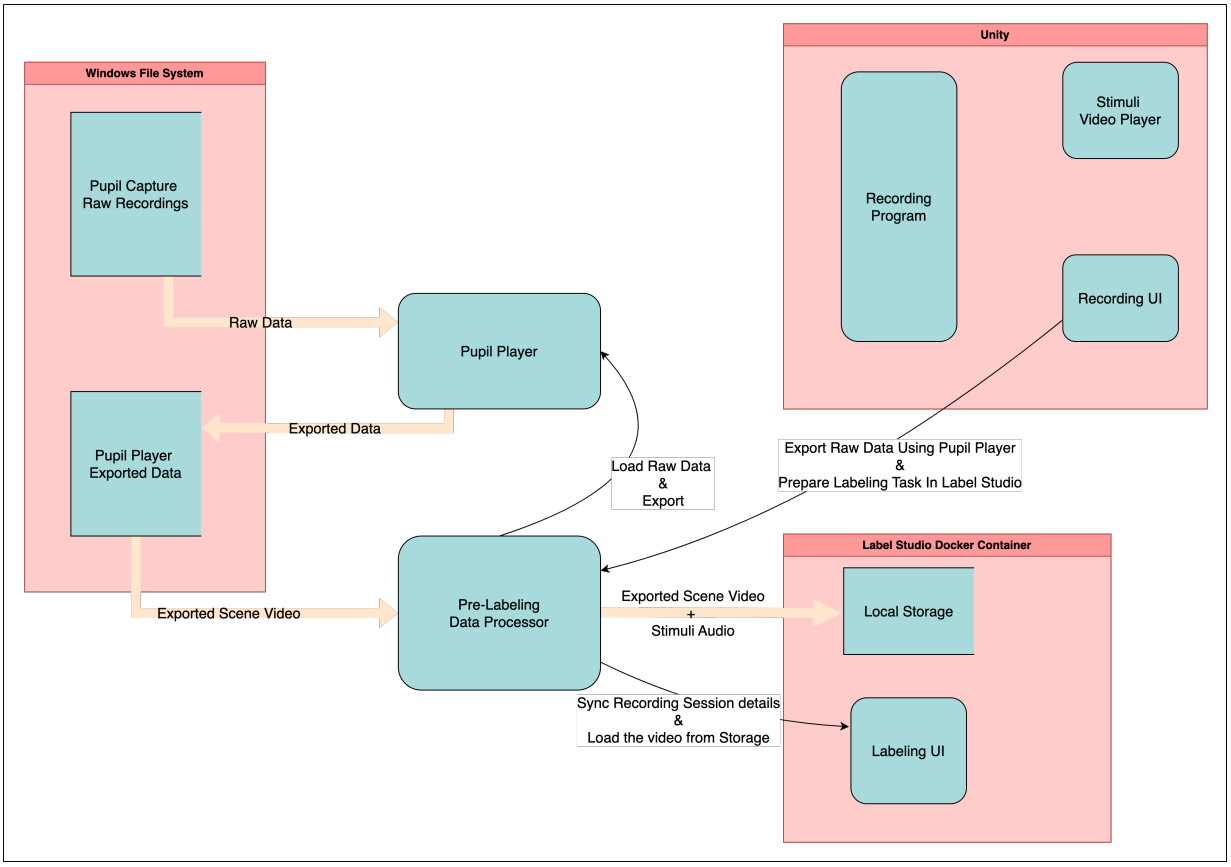


Figure 3.11: Data flow managed by the Pre-Labeling Data Processor. Pre-Labeling Data Processor is responsible for converting the raw data into a suitable format for the Labeling UI and preparing the UI for annotation tasks.

Requirements

1. Must automatically open and load recorded data into Pupil Player.
2. Must automate the conversion of raw recorded data into general-purpose formats (.mp4, .csv) through Pupil Player.
3. Must automatically add an audio track, extracted from the stimuli video, to the exported scene videos and transfer them into Label Studio's local storage.
4. Must create a new project in Label Studio with the session's name from Unity if it doesn't exist.
5. Must add new annotation tasks to the Label Studio project, loading exported scene videos into the labeling interface.

Design Considerations

The pre-labeling data processing should be automated to minimize the potential for inconsistencies and errors that may arise from manual operations. The operations involve handling and transferring data, as well as opening various external tools and executing multiple actions including text inputs within them. Making the process automatic also enhances the speed and efficiency of data collection.

Label Studio requires the videos that are to be annotated to contain audio. However, Pupil Player exports all videos without audio. Thus, an audio track must be added to the exported scene video to make it usable with Label Studio. One possibility is to add the audio recorded from the participant. But, since these recordings are mostly silent, adding the original audio track from the stimuli video would be better because the sound effects in the original stimuli videos play an important role in emotion elicitation. Thus, being able to hear them helps the participants recall their emotional responses better when reviewing the videos.

Implementation
In order to meet the first and second requirements of automatically opening Pupil Player, loading the recorded data, and exporting them, a process is started within Unity which executes a Python script. The Python script is responsible for activating the Python virtual environment (venv) required for running Pupil Player from source code, loading the raw data, initiating the 'Export' operation and waiting until it is done, and finally terminating the Pupil Player window. Pupil Player does not provide an API to perform these operations, so this had to be implemented using Windows shell commands run through a Python subprocess.

To meet the third requirement of extracting audio track from stimuli video, adding it to the exported scene video, and transferring the scene video with the added audio track to the local storage of Label Studio, another process is started within Unity to execute a Python script. This script utilizes FFmpeg, an open-source multimedia handling software, to extract the audio track and merge it with the scene video through its command-line instructions. It then establishes a connection with Label Studio via the Label Studio SDK. Using the Label Studio API, it verifies the Local Storage directory is connected as a source for importing files into the project and uploads the scene video. It then checks for an existing project in Label Studio named after the recording session in Unity. If the project doesn't exist, it creates a new project, configures its labeling interface, imports the modified scene video, and generates a new annotation task. Hereby also satisfying the fourth and fifth requirements.

3.2.9. Post-Labeling Data Processor

After labeling, this component performs additional processing such as filtering, data reorganization and integration of the labels with the collected data to form the final labeled dataset. The data flow that the Post-Labeling Data Processor must manage is illustrated in Figure 3.11.

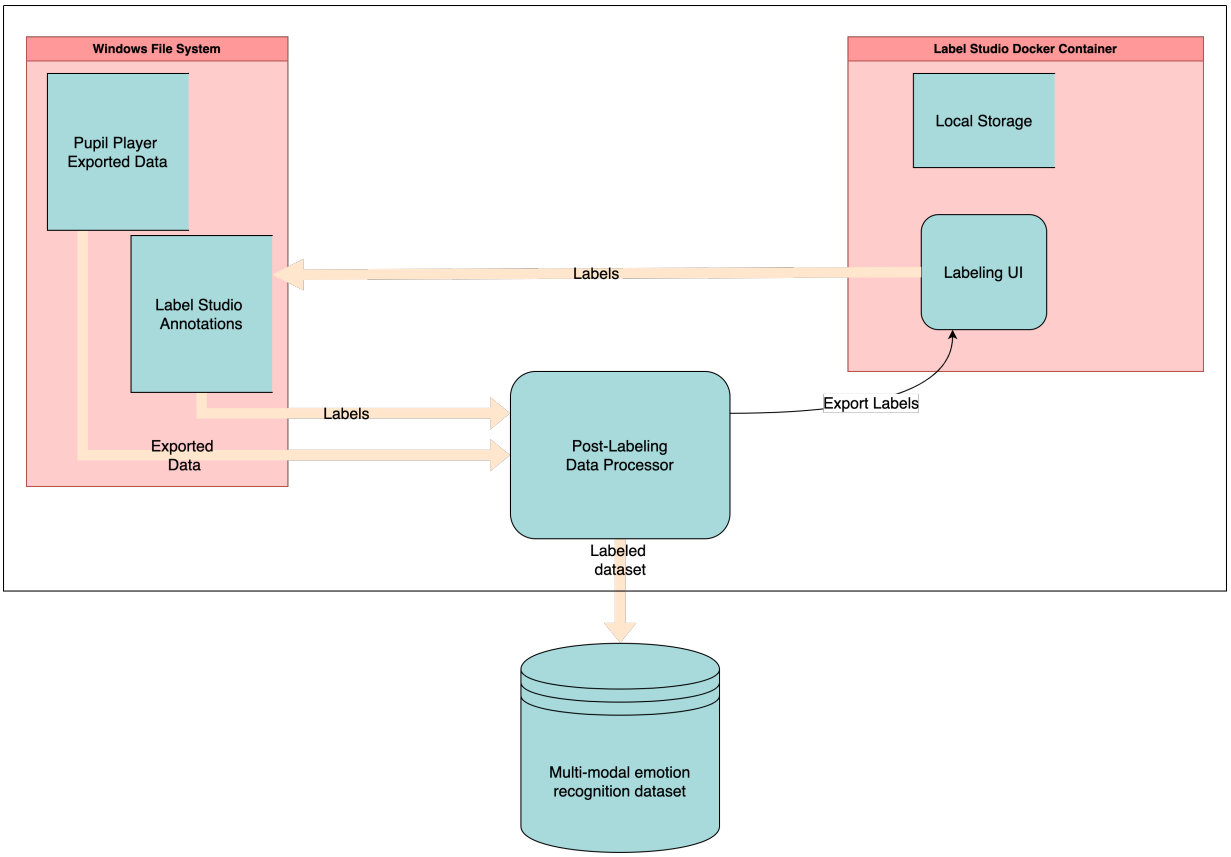


Figure 3.12: Data flow managed by the Post-Labeling Data Processor. The Post-Labeling Data Processor is responsible for processing the recorded data and labels and compiling the final dataset.

Usage Scenario

1. The Researcher runs the Post-Labeling Data Processor script.
2. The Post-Labeling Data Processor extracts the start time, end time, emotion class and emotion intensity ratings from all annotated segments
3. The Post-Labeling Data Processor distributes the annotations to all data modalities.

Requirements

1. Must export annotations from Label Studio.
2. Must extract the start time, end time, emotion class and emotion intensity ratings from all segments in all the annotated stimuli videos.
3. Must make sure the annotated segments are duplicated in all data modalities.

Design Considerations

The structure of the dataset is illustrated in Figure 3.13. For each participant, there are 14 stimuli video directories and for each stimuli video, there are 7 directories that contain the different data modalities. Labeling each data point across all modalities would thus not be a trivial task. However, since all the data entries contain a timestamp, this problem is largely simplified because it is actually not necessary to label each data point. Each exported annotation contains the start and end times of the segments in the stimuli video timeline. Thus by just taking the timestamps of the stimuli video as the reference and adjusting the timestamps of all other modalities, the distribution of the annotations over the whole dataset can be accomplished.

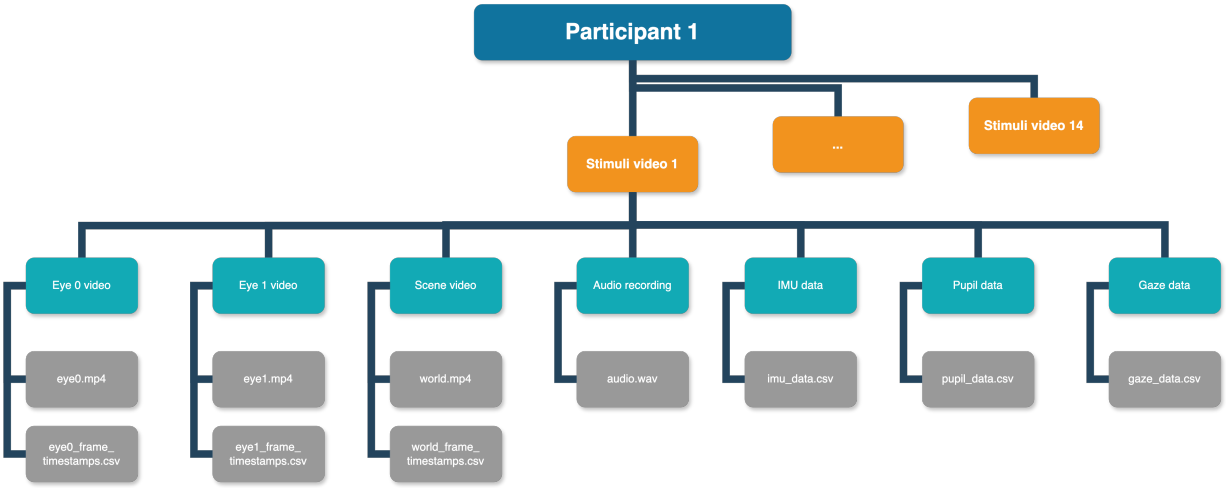


Figure 3.13: The structure of the Dataset. The Dataset contains separate directories for each participant. This figure illustrates the structure of one of the participants. The data for each stimuli video is stored separately within the participant's directory.

Implementation

The implementation of the Post-Labeling Data Processor is done in a Python Script. To fulfill the first requirement of exporting the annotations from Label Studio, the script uses the Label Studio API to connect with Label Studio and issue the Export command. Since the exported annotations contain a lot of information not relevant to our needs, only the segment details are extracted. These details include the segment start and end times, the stimuli video name, emotion class and emotion intensity rating, hereby fulfilling the second requirement.

Since the timestamps for data points across all modalities are directly recorded from the Windows System time, they do not begin at 0 seconds. To ensure alignment with the segments extracted from the stimuli video timeline, the timestamp of the stimuli video's first frame is used as a zero reference point. Its value is then subtracted from the timestamps of all other modalities. With this adjustment, the timestamps of all modalities are synchronized with the stimuli video, making the segment start and end times applicable across all modalities, thus meeting the third requirement.

3.3. Data collection environment

The physical space in which the data collection has been done can be seen in Figure 3.14. It was a quiet and isolated room, away from external disturbances such as public areas and loud machinery to ensure that the audio recording, which is one of the modalities, is not interfered. The space was provided with an ergonomic chair and comfortable room temperature to ensure the participants felt comfortable to help participants relax, which leads to more natural emotional responses. Also, the participants were asked to turn off their phones to minimize distractions and ensure that their emotional reactions were in response to the intended stimuli only.

Figure 3.14: The physical space where the data was collected.

4

The Labeling Process

In this chapter, different approaches for establishing the ground truth for the dataset are compared and evaluated, and the encountered challenges are addressed. The ground truth refers to the actual emotional experience of a participant as it occurs in response to the stimuli. Setting accurate ground truth is one of the most important aspects of the creation of a dataset because, without correct ground truth, any subsequent analysis, modeling or interpretation would be flawed, leading to incorrect research findings and conclusions.

Finding ground truths for emotions is challenging because emotions are inherently subjective, with significant variability in how individuals experience and express the same emotion. Self-reporting is mostly the method used for setting ground truth in Emotion Recognition studies [29]. Although emotional responses can be unconscious or subconscious, thus not always reaching the level of awareness necessary for self-reporting, the participant's self-reported experience is the closest approximation to accessing the internal emotional state of the individual.

4.1. Continuous Annotation

Applying a single emotion label to an entire video session would be insufficient because a single stimuli video can evoke a range of emotions at different points in time. A stimuli video session usually comprises multiple scenes or stimuli, each potentially evoking a different type and intensity of emotion. By assigning one label to the whole video, these temporal dynamics are not properly represented, which results in a loss of valuable information and can lead to misrepresentations of the emotional content of a video session. It is thus important to capture these temporal dynamics by labeling the individual scenes or segments within the stimuli videos.

A possible way to capture the temporal dynamics of emotional responses throughout the video is by having the participants actively annotate their emotional experiences in real-time as they watch a video. This method is called continuous annotation and is used by Xue et. al. when creating the CEAP-360VR dataset [17]. In their study, participants were provided with a joystick to perform the annotation. The joystick's two axes represented valence and arousal and the displacement of the joystick from its central position quantified the intensity of the emotional response. It is important to note that, unlike the discrete basic emotions used in this study, Xue et. al. opted for the Circumplex model of emotion which only indicates the valence and arousal levels. Also, the stimuli videos that they used were mostly of landscapes and activities which demanded minimal cognitive effort from participants, thus making it easier for participants to continuously annotate their emotions.

4.2. Simplifying Continuous Annotation: From Joystick to Trigger and Button

Using joysticks for continuous annotation would not be a viable option for this study because it would require the participants to use two joysticks and there would be no straightforward way to remember which axes of the joysticks represent which emotions, thus significantly increasing the cognitive load. An attempt

to simplify the method was made by using a trigger mechanism instead of joysticks. Since discrete emotion classes are used in this study and the stimuli videos are designed to evoke one emotion, only the intensity of the emotional response would need to be annotated. To record only the intensity, a trigger present in the VR controller would be sufficient. However, during preliminary testing, it was quickly found that having the participant press the trigger to quantify the intensity proved to be too distracting. The stimuli videos chosen for this study are all extracted from movies, which contain dialogues and actions that require attention from the participants to be effective. If the participants were to constantly think of how intense they are experiencing an emotion and how much to press the trigger, it would put a significant cognitive load on them and they would not be immersed in the content of the video, making the stimuli video less effective at evoking the intended emotion.

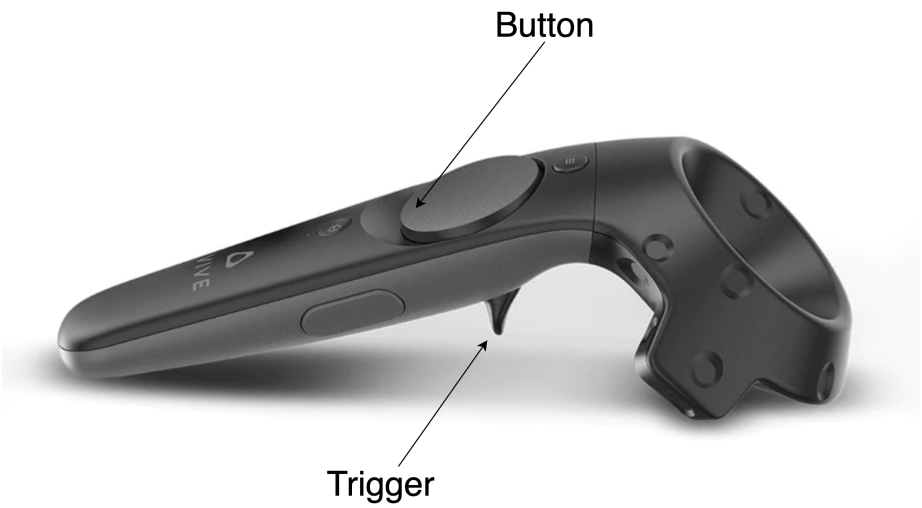


Figure 4.1: The trigger and the button on the VIVE Pro VR controller.

Using the trigger mechanism required the participant to make two decisions: whether to press the trigger and to what extent. To reduce the cognitive load for the participants, this approach was revised to have them press the trigger only to mark the segments when they felt an emotion strongly enough so that they could assign the intensity for each marked segment right after watching the video. Also, pressing the trigger on the controller required more physical exertion than pressing a button, so it was decided to mark the segments using a button. However, even this proved to be too distracting because the participants had to constantly think whether the emotion they were feeling was strong enough to press the button. Just having to constantly think whether or not to press the button distracted the participant from properly following the videos which required their full attention to effectively evoke the intended emotions, so this method was also dropped.

4.3. Experimenting with Semi-Automatic Labeling using Deep Learning

To maximize the effectiveness of the stimuli videos by minimizing all external cognitive load on the participants, allowing them to be fully immersed in the content of the video, it was decided to do the annotation right after the completion of each video. However, recalling all the segments in which the participants clearly felt an emotion was too difficult for the participants. Thus taking inspiration from the work of Zhao et. al. in the making of the EMOShip-Net, where they used a Binary Classifier to detect Non-neutral emotions in real-time, it was decided to use the same binary classifier to mark the non-neutral

segments in the video automatically. The workflow of the EMOShip-Net is shown in Figure 4.2. The EMOShip-Net receives and analyzes two video streams: one from the Eye Camera and the other from the Scene Camera. If a non-neutral emotional state is detected through the Eye Camera stream by the Binary Classifier, it triggers the rest of the system to further analyze and classify the non-neutral emotion by examining the corresponding scene captured at that moment.

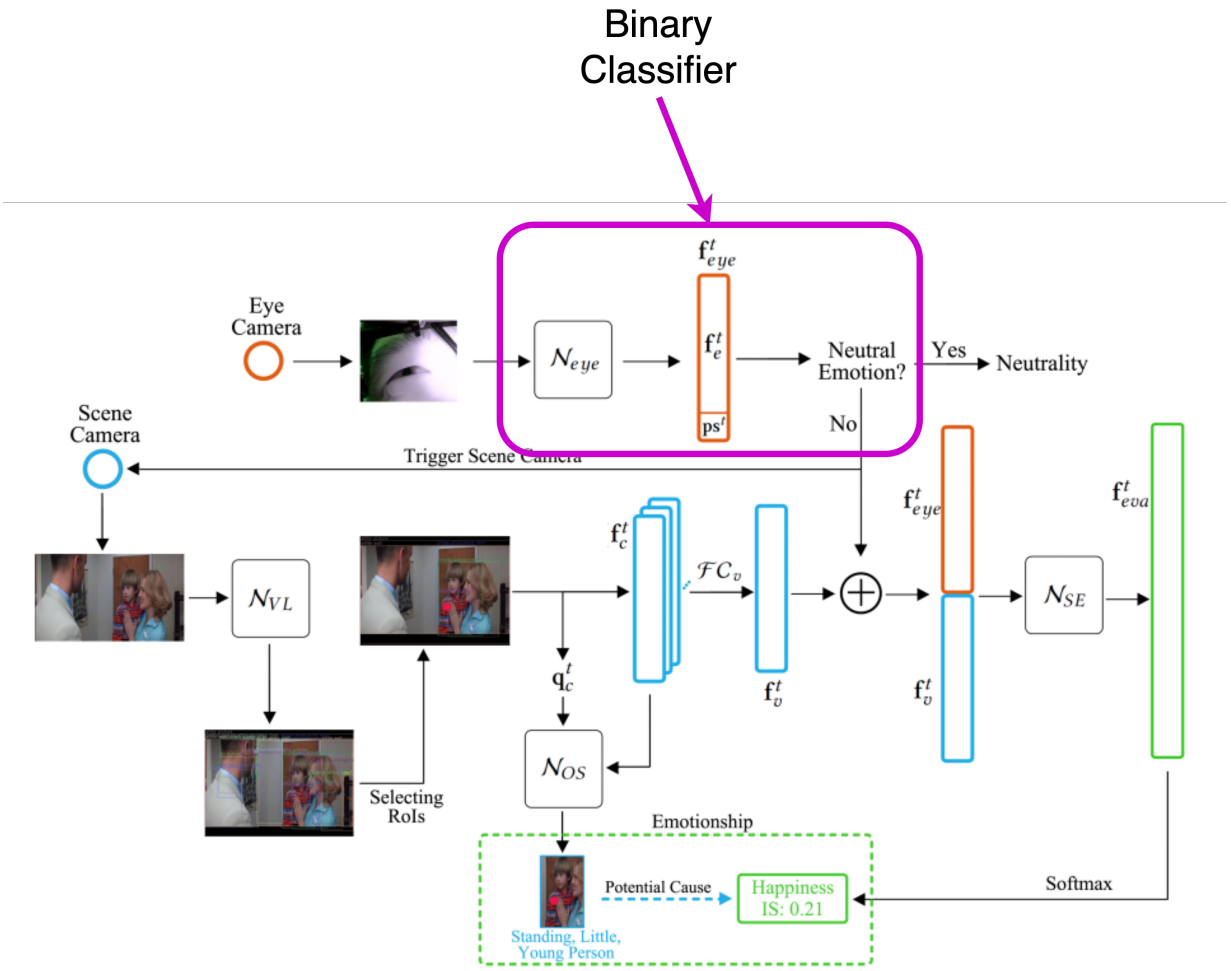


Figure 4.2: Binary Classifier in the workflow of EMOShip-Net [16]. In this model, the binary classifier is used as a trigger to run the full classification process only when non-neutral emotion is detected, as an energy-saving mechanism.

The idea was to use the same Binary Classifier to analyze the participants' eye images after they had watched a stimuli video and automatically mark segments on the video timeline where non-neutral emotional states were detected. After the non-neutral segments are marked by the Binary Classifier, the participant would then only need to give those marked segments a numerical rating representing the experienced emotional intensity. The Binary Classifier in the EMOShip-Net was implemented using a modified ResNet-18 network to extract features from the eye images and a binary classification layer was added to distinguish between neutral and non-neutral emotional states.

Hypothesis

When a participant watches a stimuli video, their default baseline emotional state should mostly be Neutral and there should be a few segments in the video where Non-Neutral emotional responses are evoked. It is assumed that the facial expressions associated with Neutral state are uniform thus these expressions can be grouped together using a clustering method. A pre-trained ResNet-18 network can be used to extract features from the eye images and a clustering method can be used to cluster these

features into Neutral and Non-Neutral clusters. Since the stimuli videos are designed to evoke one specific emotion, the two main clusters are anticipated to correspond to the Neutral emotional state and the specific Non-Neutral emotional state targeted by the video. Thus by aligning the eye image frames, from which the clustered features are extracted, in their proper sequential order, segments in the video timeline can be marked corresponding to the Neutral and Non-Neutral emotional states.

Implementation

A pre-trained ResNet-18 model, taken from the torchvision.models library, was further trained using the dataset provided by Zhao et. al., which consisted of eye images with 7 emotion labels. This was done to improve the network's ability to extract relevant features from the eye images. After the fine-tuning, the model's final classification layer was removed to retain only the feature extraction component of the network.

The feature extractor is then used to process a recorded eye video frame-by-frame to generate a corresponding feature vector for every frame. The feature vectors are then grouped into two clusters using an agglomerative hierarchical clustering method. After assigning a cluster number to each of the feature vectors, they are then reordered to their original sequence.

Testing

In order to test the Binary Classifier, an eye video recorded during the presentation of a Surprise stimulus was selected. For this particular video, the test participant rated their level of surprise as 9. The stimulus video primarily contains neutral scenes with a surprise-evoking scene shown right at the end. This video was assumed to be an optimal test case for the Binary Classifier because it mostly contains neutral segments which should elicit a uniform facial expression and only one segment which should elicit a distinct expression.

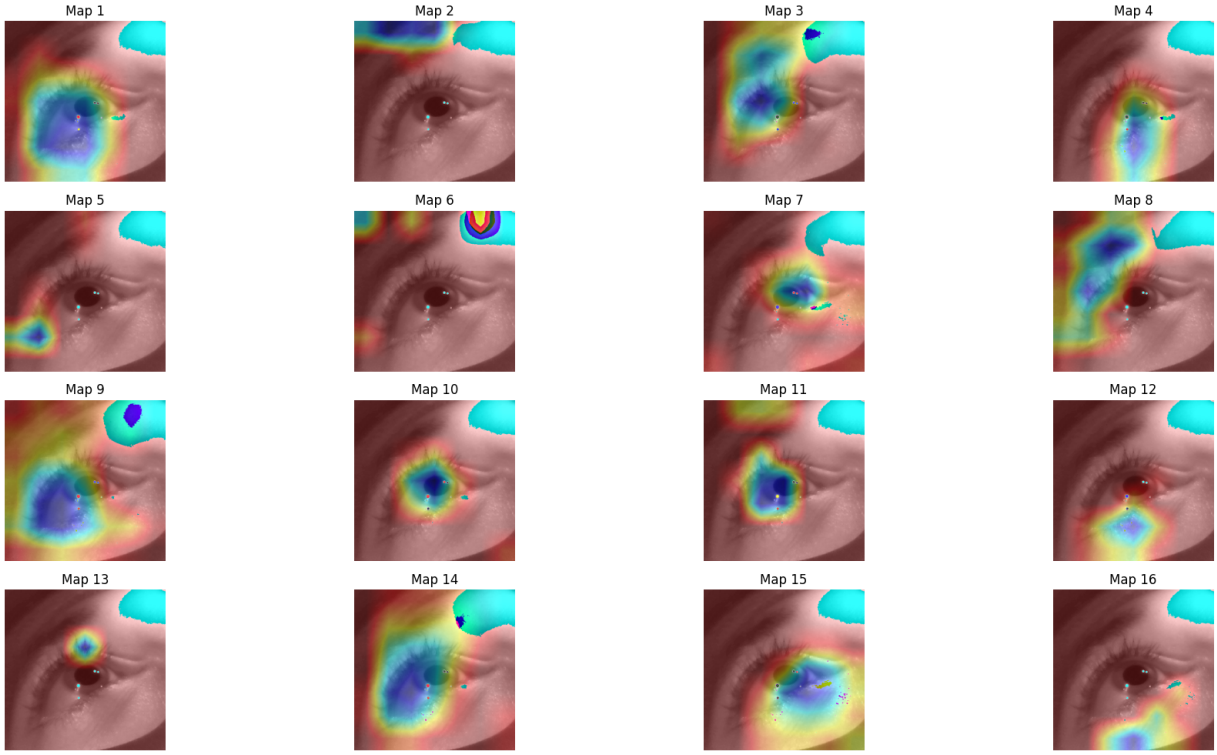


Figure 4.3: Visualization of some of the features extracted using ResNet-18.

First, as a preliminary test of the feature extraction process, the first 16 out of 512 feature maps were visualized by overlaying them as heat maps on a frame from the video, as shown in Figure 4.3. Each map represents the spatial activation of different filters in the last convolutional layer of the ResNet-18 model. It can be observed that the extracted features mostly come from areas such as the iris (maps 7, 10, 11), the

sclera (maps 1,9,14), and the eyebrows (maps 2,3,8). This suggests that the feature extractor is effectively extracting the relevant features.

Results

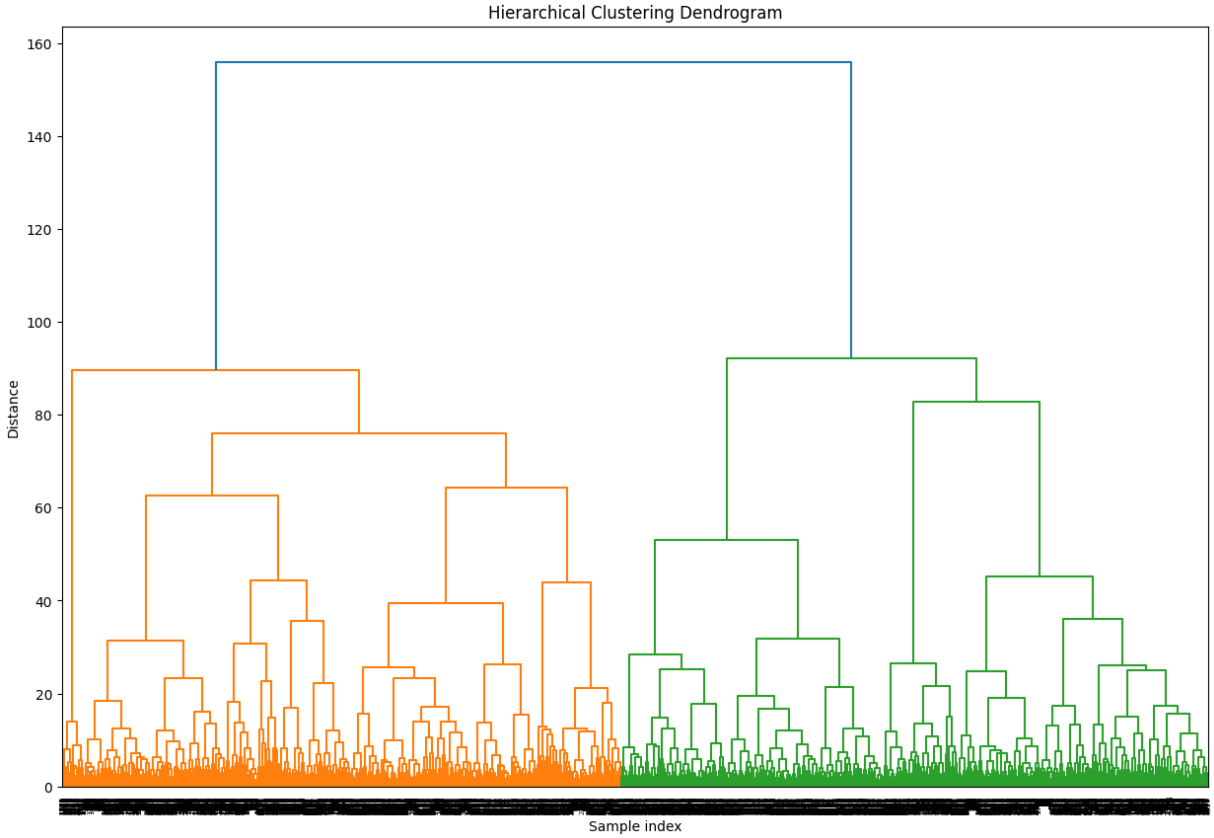


Figure 4.4: Dendrogram showing hierarchical clustering of the feature vectors corresponding to the eye video frames.

It can be observed in the dendrogram in Figure 4.4 that two distinct clusters have been found. However, both the cluster are almost equal in size, which is not as expected. Since the surprise video contains mostly neutral scenes and only a surprise evoking scene towards the end, it was expected that there would be one big cluster representing the neutral emotion and a smaller cluster representing the surprise emotion. In Figure 4.5, the cluster assignments of all the frames in their original temporal order is plotted. This plot confirms that the surprise segment of the video was not correctly detected. The experiment was repeated using different linkage types and also different clustering methods including K-Means and HDBSCAN methods. However, the results were similar.

Conclusion

The hypothesis was proven wrong. The Binary Classifier for Non-neutral segments detection, which was inspired by the EMOship-Net's Binary Classifer (Figure 4.2), was not able to detect the Non-Neutral segments using a ResNET-18 model. The authors of the EMOship-Net do not analyze and evaluate the performance of their Binary Classifer separately, thus it was not possible to make a direct comparision. They used the Binary Classifier only as a energy-saving mechanism in their main model, thus its proper working was not crucial for the accuracy of the overall model.

Further analysis of this Binary Classifier was not done because the assumption that a ResNET-18 model alone would be able to detect emotional patterns may have been overly optimistic. Emotional responses are often subtle and have an irregular temporal aspect, thus a more complex and sophisticated model would be necessary to reliably detect them. And the primary goal of this project is to create a dataset to enable the development of such complex models, thus a circular problem was encountered. So, other

options have to be explored to solve the problem of automatically marking non-neutral segments or help participants recall the segments in which they clearly felt an emotion.

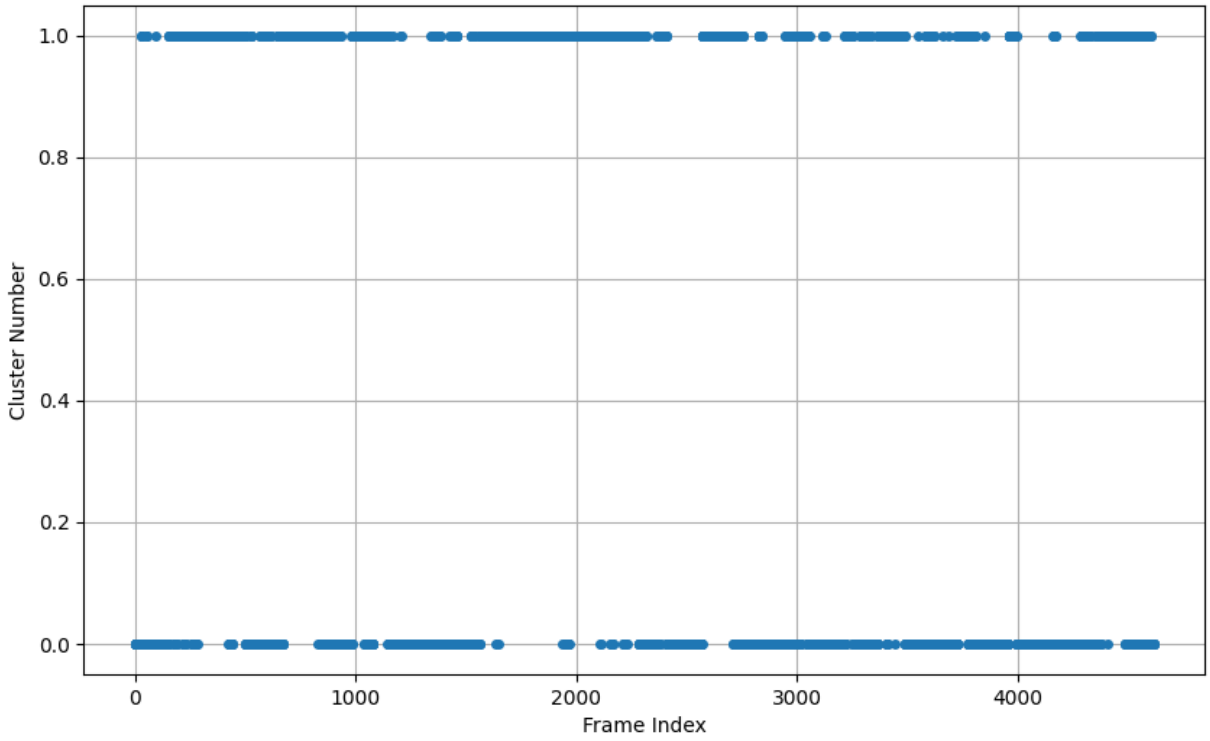


Figure 4.5: Cluster assignment for each frame of the recorded eye video.

4.4. Final: Assisted Labeling with Manually Created Segments

The aim of the Binary Classifier was to help the participants recall the segments where they experienced clear emotional responses after watching the stimuli video by automatically predicting and marking the segments. Because continuous annotation by participants while watching the stimuli was not an option as it would disrupt their immersion in the video, making the stimuli less effective. A simple approach to solving the problem would be by manually predicting and marking the segments. Since the stimuli videos contain different scenes, the researcher could analyze each scene for emotional triggers and predict which scenes/segments would be most likely to evoke the intended emotions. The predicted segments could then be presented to the participants during the labeling process so they can easily recall if the segments were effective at evoking the intended emotion.

Following this approach, the following labeling process was created :

1. The participant watches the stimuli video without any distractions.
2. After the video has finished, the recorded data is pre-processed and prepared for labeling.
3. During the pre-processing, the manually created segments for the stimuli video are added to the video timeline as suggested non-neutral segments by the Pre-Labeling Data Processor Component.
4. The participant can easily jump to the suggested segments and replay them individually.
5. The participant can choose to modify the segments' length and position on the video timeline to more accurately define them.
6. The participant can choose to create new segments on the timeline if a segment they clearly remember is not suggested.
7. For each segment, the participant assigns an emotion class.
8. For each segment, the participant gives a numerical intensity rating.

5

Dataset Validation

The validation of an emotion recognition dataset is a challenging task because a direct measurement of internal emotional experiences is not feasible with current technology. Emotions are nuanced and subjective as they are expressed slightly differently by different persons, which makes it difficult to manually identify clear indicators of a particular emotional response. And the expression is done through a combination of various modalities, mostly involuntarily and subconsciously or unconsciously. Especially the modalities that are captured using a VR headset, which mostly include subtle involuntary movements of the eyes, the regions around the eyes and the head. Recognizing a pattern for each emotion with these subtle movements cannot be done manually. Humans possess the natural ability to detect and interpret these subtle movements and associate them with specific emotions subconsciously. To consciously detect these movements would require advanced emotion recognition systems. And the aim of the creation of this dataset is to help develop such advanced systems, thus this presents a circular problem.

The most feasible approach to verify that the dataset contains genuine emotional response indicators currently is to evaluate the self-reports of the participants. Even though they are subjective, they are the closest available representations of the ground truth. Since the modalities captured using VR contain subtle involuntary and thus objective responses, it would be the function of the future deep-learning emotion recognition systems to detect and identify these patterns. By providing the annotated segments in the data, the development of these systems is assisted by narrowing down the parts in the data which most likely contain the objective emotional response patterns.

Still, an attempt has been made to verify that the data actually contains usable information related to the emotional responses by analyzing one of the modalities. Pupil responses are involuntary and are known to be clear indicators of emotional arousal. Although, the pupils themselves do not say anything about the emotions, they do indicate emotional arousal levels. And since Russell's circumflex model of emotions indicates that the discrete emotions can be placed along the valence-arousal dimensions, is should be possible to see some difference in pupil responses between the emotions, especially between the contrasting emotions, which are sadness and happiness or surprise.

The validation of the dataset has thus been decided to do in two steps. Firstly, the participants' self-reported emotional intensity ratings are evaluated to confirm that the data collection process was effective at evoking the intended emotions effectively, and thus the responses are captured in the recorded data. And secondly, the pupil diameter data is processed and analyzed to check if it can be indicative of the different emotions.

5.1. Participant Self-Reported Ratings

The pool of participants primarily consists of students and staff from the university. To ensure the reliability of the collected data, the participants have been screened for visual and auditory impairments and other medical conditions such as heart conditions, claustrophobia, motion sickness, or epilepsy/seizure disorders that could influence their emotional state. Also, it has been attempted to ensure the participant demographics were as diverse as possible, as illustrated in Figures 5.1, 5.2, and 5.3, showing participants' age, biological sex, and ethnic background, respectively.

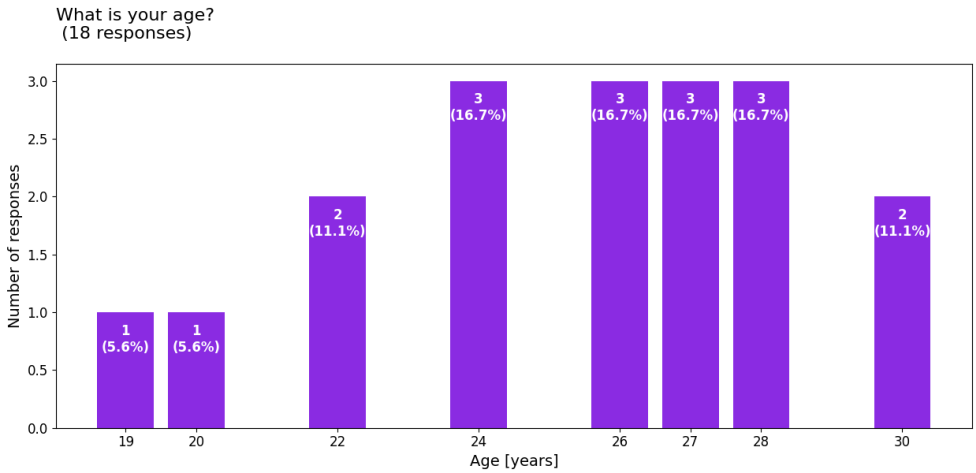


Figure 5.1: Participants' age

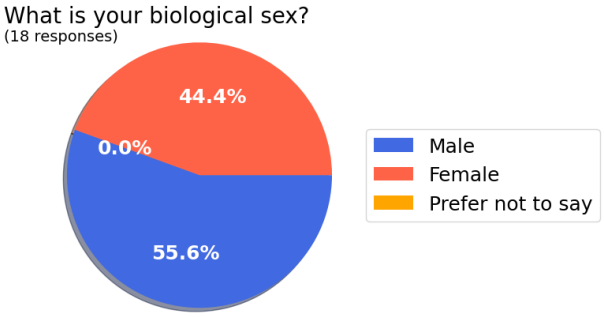


Figure 5.2: Participants' biological sex

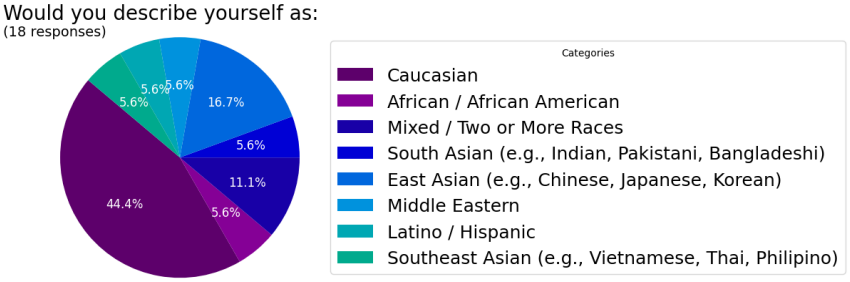


Figure 5.3: Participants' ethnic background

The self-reported ratings for the intensity of the experienced emotions during the defined segments in the videos are summarized in Figure 5.4. The values depicted in blue within the boxplots represent the median rating for all the segments belonging to the same emotion. The mean and standard deviation values of the ratings are displayed on the right side of the plot in purple. In green, the percentages of the segments belonging to the same emotion that were rated higher than 6.0 are displayed. The participants were asked to rate each segment on a scale of 1 to 10. The question that was posed to them was: "How happy/sad/angry/etc. would you say you felt during this section of the video on a scale of 1 to 10?". A rating of 1 indicated that they did not experience the emotion at all and 10 indicated that they experienced the emotion very intensely. A rating of 6 generally indicated that there was a clear experience of the emotion but the intensity was close to minimum. The median rating for the emotions of Sadness and Disgust is 8.0, and for Surprise, Happiness, Anger, and Fear is 7.0. Also, at least 70% of the segments were rated above

6.0 for all the emotions, which suggests that the stimuli were quite effective at evoking all the intended emotions.

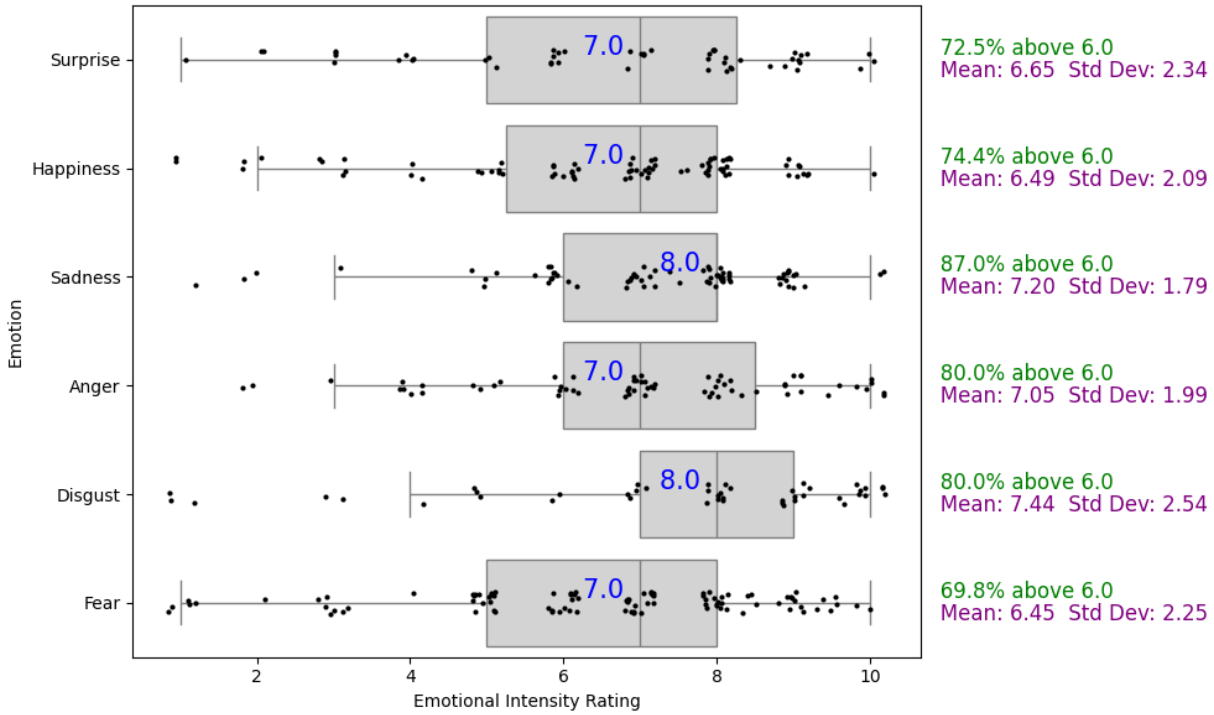


Figure 5.4: Boxplot with Scatter of Emotional Intensity Ratings by Emotion. The blue number within the boxplots denotes the median value of the ratings for all the segments belonging to the same emotion.

5.2. Pupil Diameter Analysis

In the following sections, the step-by-step analysis of the recorded pupil diameter data is described. First, the data is pre-processed to remove blinks, noise and outliers. Then, the problem of the luminance influence on pupil diameter is addressed to isolate only the effect of emotions on the pupil diameter. The isolated pupil diameter data corresponding to emotions only is then qualitatively and quantitatively analyzed to confirm the presence of emotional responses with regards to the presented stimuli for different emotions.

5.2.1. Pre-processing Pupil Data

In figure 5.5, the raw pupil diameter data of left and right eye, along with the luminance level of the video is shown. The data is filtered only using the confidence level, provided by the Pupil Labs software. Pupil Labs suggests that only the data above confidence level 0.6 is useful, thus this is chosen as the filtration threshold.

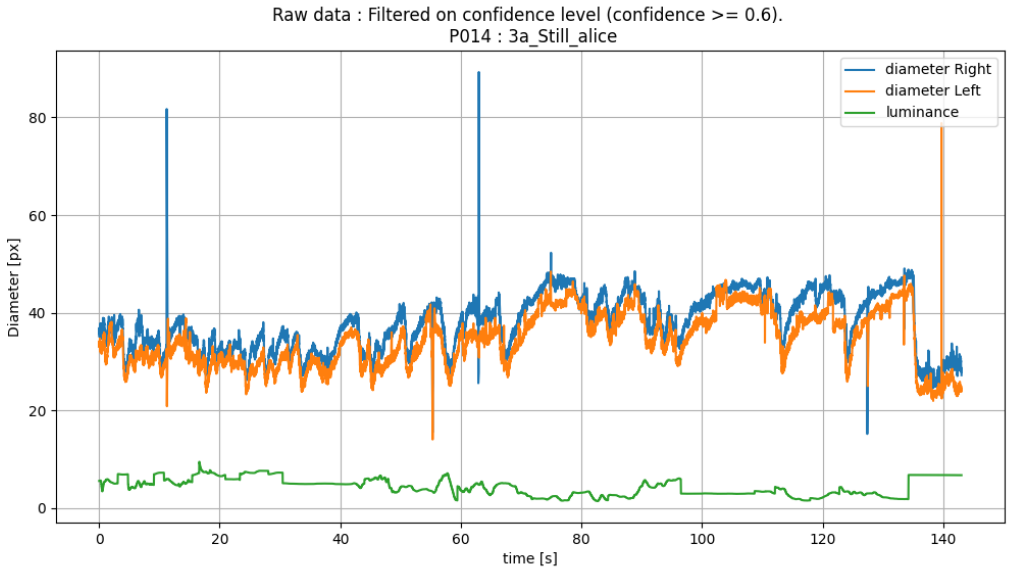


Figure 5.5: Left and Right pupil diameters. Raw, only filtered on the confidence level.

In the raw data, some significant outlying peaks can be observed. These peaks are caused due to false pupil detection by the Pupil Labs software which sometimes occurs during blinks. The software sometimes falsely detects a pupil with a confidence level higher than 0.6, which results in these peaks getting through the initial simple filtering based on the confidence level. To remove these, a more advanced outlier detection and filtering process needs to be used.

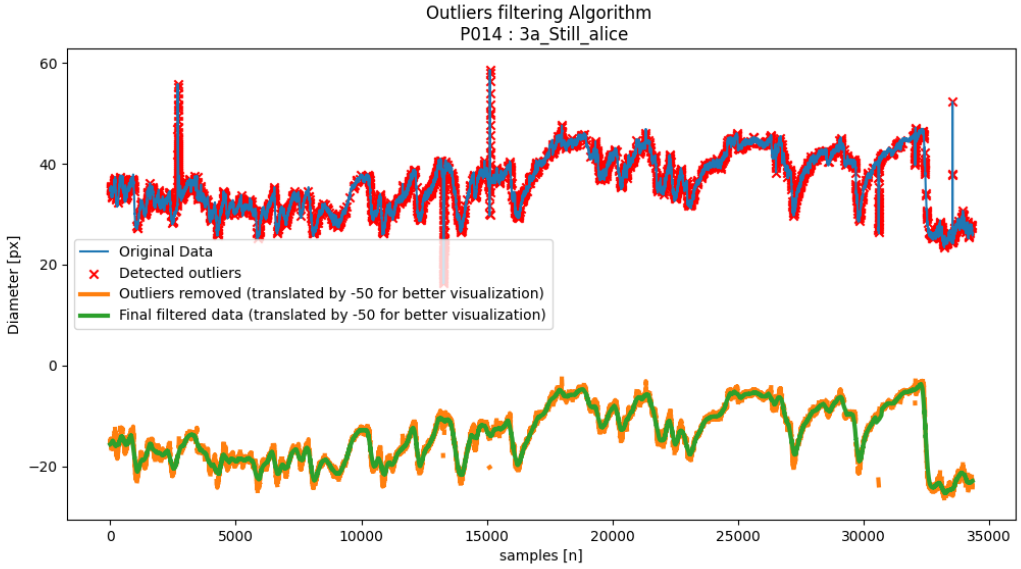


Figure 5.6: Outliers filtering algorithm : intermediate results

In figure 5.6, the intermediate results of the outliers filtration algorithm has been shown. Firstly, an average of the left and right pupil diameter data is taken and re-sampled at 240Hz. This is represented by the blue line called 'Original Data' in the figure. The raw data was recorded at around 120Hz (targeted) but a consistent sampling frequency was not guaranteed by Pupil Labs software. Thus, for easier processing, the recorded data has been re-sampled at 240Hz with linear interpolation.

Secondly, the derivative of the original data is taken, which represents the change/speed of the pupil diameter over time. Since the pupil can dilate and constrict at a certain maximum speed, this measure is more useful for outlier detection than the actual pupil diameter, which can be very different for different people and different ambient illumination levels. Thirdly, the Median Absolute Deviation (MAD) of the derivative is calculated and used as the threshold for outlier detection. MAD has been chosen because it is robust against outliers since it uses the median and is thus less affected by outliers than, for example, mean or standard deviation which are heavily influenced by the extreme values of the outliers. The detected outliers in the original data have been marked by red 'x' markers in figure 5.6. And, the data with outliers filtered is shown by the orange graph in the same figure. The graph has been translated by -50 along the y-axis for better visualization.

Fourthly, a sliding window local outlier Detection and Replacement based on z-score has been used to smoothen the filtered data. This was necessary because, when using the MAD of the derivative outlier filtration, some residue of the outlying peaks remains as it can be seen in the orange graph at sample number 15000 in figure 5.6. By simply linearly interpolating the outlier filtered data, the removed peaks would have been re-introduced. In this step, using the z-score, which is based on mean and standard deviation, does not pose a big problem because the MAD of the derivative approach has filtered out the extreme outliers and the remaining outliers are small in number and magnitude, thus, don't heavily influence the mean and standard deviation. The result of this step is represented by the green graph in figure 5.6 and the comparison between the raw data and the final outliers filtered data is shown in figure 5.7.

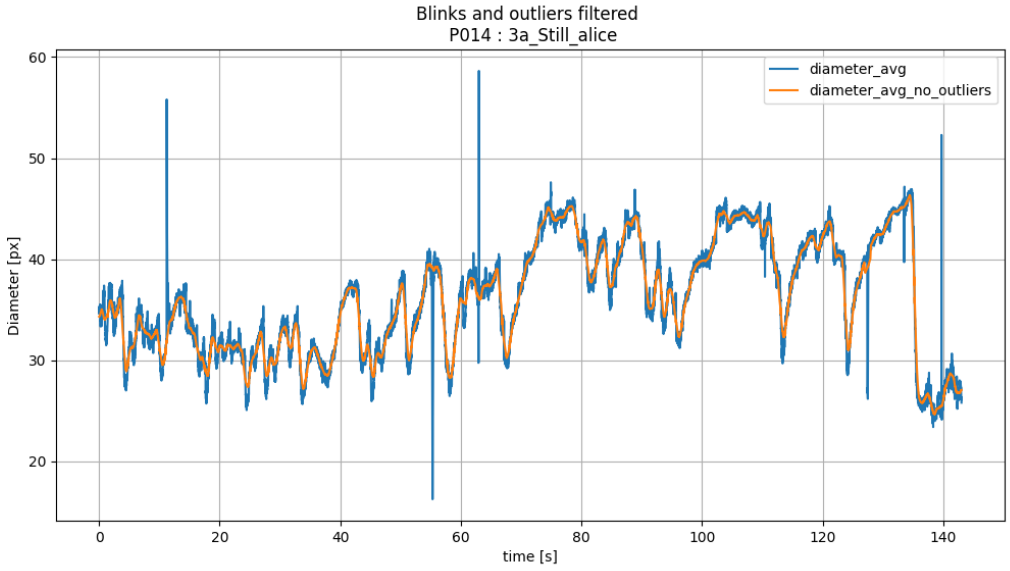


Figure 5.7: Final outliers filtered data. Average of left and right pupils.

5.2.2. Estimating the Influence of Ambient Luminance on pupil diameter

Pupil diameter not only reflects the emotional state of a person but is also heavily influenced by the ambient illumination levels [30]. The pupil constricts for higher illumination levels and dilates for lower illumination levels. Since the pupil diameter data is recorded from participants wearing a VR HMD, all the ambient light influencing the pupil diameter comes from the HMD when playing the videos. The illumination levels can be retrieved for each frame of the videos by converting the frames of the videos into HSV color format and isolating the 'V' component which represents the luminance level. The luminance graph of one of the videos is shown and represented by the green line in figure 5.5.

Since the luminance comes from only one known source, its effect on the pupil diameter can be modeled by equation 5.1 [17]. In the equation, PD represents the recorded pupil diameter, $PD_{luminance}$ represents

the influence of luminance on the pupil diameter and $PD_{emotion}$ represents the influence of the participant's emotional/mental state on pupil diameter.

$$PD = PD_{\text{luminance}} + PD_{\text{emotion}} \tag{5.1}$$

The luminance component of the pupil diameter can be modeled by equation 5.2, where LV represents the luminance level value, a represents the scaling factor and b represents the offset. This is because the relationship between ambient luminance level and the pupil diameter is linear [31].

$$PD_{luminance} \approx PD_{luminance\_est} = a \times LV + b \tag{5.2}$$

Following the work of Xue et. al. on a similar study [17], a linear regression model has been used to estimate the luminance component of the pupil diameter ($PD_{\text{luminance\_est}}$). Note that, the luminance level is inversely correlated with pupil diameter, which is why the estimated graph is seen flipped in figure 5.8.

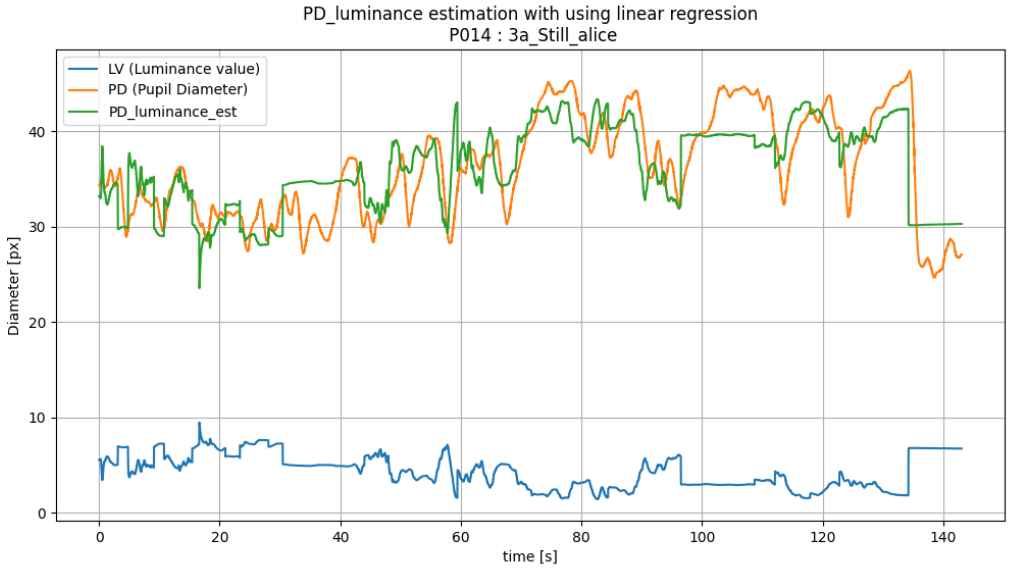


Figure 5.8: Estimation of the luminance component of the pupil diameter ($PD_{\text{luminance\_est}}$) using Linear Regression

One notable difference in the behavior between $PD_{\text{luminance\_est}}$ and PD (pupil diameter) is that $PD_{\text{luminance\_est}}$ can abruptly jump from a lower value to a higher value or higher value to a lower value in one frame, while pupil diameter gradually increases or decreases. This can be seen in figure 5.8. right before the 100 second mark, where the $PD_{\text{luminance\_est}}$ graph makes a steep rise upwards and the PD graph slowly follows. Also right before the 140 second mark, the $PD_{\text{luminance\_est}}$ graph makes a steep decline and the PD graph follows. It is also interesting to note that the increasing rate of the pupil diameter (dilation) is lower than its decreasing rate (constriction), which is as expected because of the physiological mechanisms and neural pathways involved in these actions [32].

Since the $PD_{\text{luminance\_est}}$ is supposed to be a component of PD according to equation 5.1, its behavior must match PD as closely as possible. To enforce this, a custom linear fitting algorithm has been implemented in order to fit the luminance graph to the Pupil Diameter graph while satisfying the linearity and maximum slope constraints. The result of this custom fitting algorithm is shown in figure 5.9, where the red line represents the improved $PD_{\text{luminance\_est}}$. The Root Square Mean Error (RSME) of $PD_{\text{luminance\_est}}$ has dropped from 3.326 to 2.606, a 25.65% decrease.

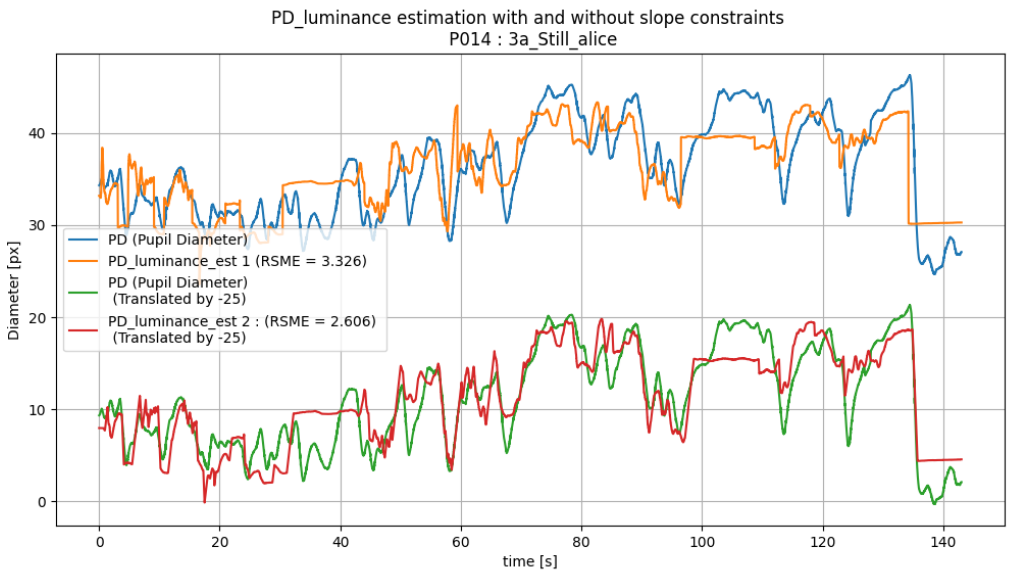


Figure 5.9: Comparison between $PD_{\text{luminance\_est}}$(1 : Linear regression) and $PD_{\text{luminance\_est}}$ (2 : Custom linear fitting algorithm with slope constraints)

Algorithm 1: Find Best Linear Fit with Grid Search

```
Input: target_graph, graph_to_fit, scaling_factor_bounds, offset_bounds, translation_bounds
Output: transformed_graph, best_scale_factor, best_offset, best_translation, min_error
max_increase_rate, max_decrease_rate ← find_max_slopes(target_graph)
for scale_factor in scaling_factors do
    for offset in offsets do
        for translation in translations do
            translated_x_values ← x_values + translation
            interpolated_graph ← np.interp(x_values, translated_x_values, graph_to_fit,
              left=graph_to_fit[0], right=graph_to_fit[-1])
            transformed_graph ← scale_factor * interpolated_graph + offset
            transformed_graph ← limit_rate_of_change(transformed_graph, max_increase_rate,
              max_decrease_rate)
            error ← np.mean((transformed_graph - target_graph) ** 2)
            if error < min_error then
                min_error ← error
                best_scale_factor ← scale_factor
                best_offset ← offset
                best_translation ← translation
            end
        end
    end
end
translated_x_values ← x_values + best_translation
interpolated_graph ← np.interp(x_values, translated_x_values, graph_to_fit, left=graph_to_fit[0],
  right=graph_to_fit[-1])
transformed_graph ← best_scale_factor * interpolated_graph + best_offset
transformed_graph ← limit_rate_of_change(transformed_graph, max_increase_rate,
  max_decrease_rate)
return transformed_graph, best_scale_factor, best_offset, best_translation, min_error
```

The custom linear fitting algorithm is presented in algorithm 1 and 2. In short, this algorithm finds the

best parameters for the linear fitting by performing a 3D grid search on the parameters: scale factor, offset on the y-axis, and translation along the x-axis. The translation along the x-axis parameter had to be added for better fitting results because the pupil reacts to changes in the luminance levels with a slight delay.

Algorithm 2: Limit Rate of Change

```
Input: signal, max_increase_rate, max_decrease_rate
Output: limited_signal
limited_signal ← np.array(signal, dtype=np.float64)
for i ← 1 to len(limited_signal) - 1 do
    increase ← limited_signal[i] - limited_signal[i - 1]
    if increase > max_increase_rate then
        limited_signal[i] ← limited_signal[i - 1] + max_increase_rate
    end
    else if increase < -max_decrease_rate then
        limited_signal[i] ← limited_signal[i - 1] - max_decrease_rate
    end
end
return limited_signal
```

5.2.3. Isolating the change in pupil diameter caused primarily by change in emotional state

In figure 5.10, the estimated component of pupil diameter caused by luminance level ($PD_{\text{luminance\_est}}$) has been subtracted from the recorded pupil diameter data to isolate the pupil diameter component caused primarily caused by the change in emotional state (PD_{emotion}) of the participant. This is represented by the green line. It can be observed that the PD_{emotion} changes quite a lot, depending on the different scenes in the video. The video clip (3a_Still_alice) was selected to evoke the emotion of sadness primarily. However, it is quite clear that not all the segments or scenes in the clip are equally sad or evoke only the sadness emotion. The green graph representing PD_{emotion} shows that there are segments where the pupil dilates (peaks), which is not associated with sadness, and there are segments where the pupil constricts (troughs), which is associated with sadness. There are also segments where the pupil diameter stays near the x-axis, which indicates neutral emotion. Thus, these segments need to be individually inspected, instead of assuming that the whole video clip evokes only one emotion.

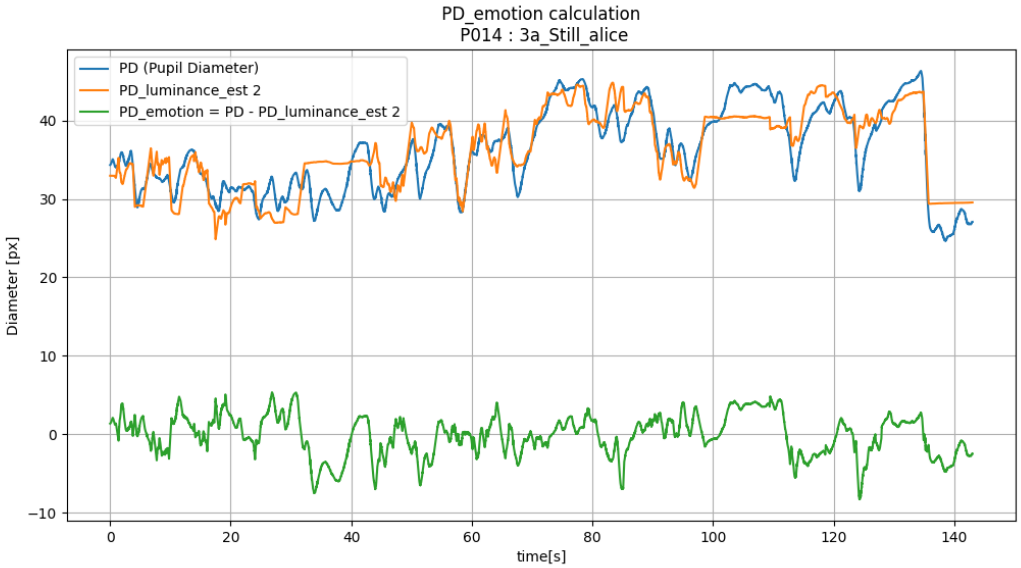


Figure 5.10: PD_{emotion} calculated by subtracting $PD_{\text{luminance\_est}}$ from PD (Pupil Diameter)

5.2.4. Inspecting the scenes in the video clips that caused the constriction (peaks) and dilation (troughs) of a participant's pupil

In figure 5.11, the peaks in the PD_{emotion} graph, which indicate pupil dilation, have been labeled with numbers 1 to 4. Also, a snapshot and a short description of the scene associated with each peak have been added. The troughs will be analyzed separately afterward.

To provide context about the video clip, the clip comes from the movie Still Alice (2014) and was selected to evoke the emotion of Sadness. In the clip, a woman and her husband are spending time together at their beach house after they find out she has early-onset Alzheimer's. Before they go for a run, the woman wants to go to the bathroom. She wanders through the house to find the bathroom, opening a series of different doors. She is unsuccessful and when her husband comes to find her she is standing in the hallway with pee in her pants and starts to cry because she cannot remember anything.

It can be seen in figure 5.11 that peaks 1,2 and 3, are caused when the woman is looking at old photos, when she is smiling at her husband who just arrived in the scene, and when the husband is shown with an amicable expression, respectively, which is as expected. However, peak 4 shows that when the woman started to almost cry after she couldn't remember where the bathroom was, the participant's pupil also dilated. This was not as expected because sadness should have caused the pupils to constrict. However, this behavior could be explained because the emotional response of each individual is subjective. In this instance, the participant was perhaps feeling more empathetic than sad. Empathy is not classified as one of the basic emotions like sadness. Empathy is more complex and actually causes pupils to dilate [33], unlike sadness, even though these emotions seem similar. It is also important to note that the woman doesn't actually cry in this segment. She cries a few moments after in another segment, which does cause the pupils to constrict as expected, indicating that the participant may have experienced sadness this time. This will be discussed in the next paragraph.

Figure 5.11: Analyzing the PD_{emotion} peaks and the corresponding scenes

In figure 5.12, the troughs in the PD_{emotion} graph, which indicate pupil constriction, have been labeled with numbers 1 to 6. Trough 1 occurs when the woman forgets that she just agreed to go running with her husband and continues to look at the photos so the husband asks her again. This gives the first glimpse of her condition. when Trough 2 occurs, the woman enters the house and as soon as she enters, her facial expression shows that she has forgotten where she was going. This gives a second glimpse of her condition. With Trough 3, the woman is clearly confused and the sad background music intensifies. With Trough 4, the camera focuses on the woman's distressed face after she hastily opens several doors to find the bathroom. With Trough 5, it is shown that the woman has peed her pants. And finally, in Trough 6, the

woman starts to cry, when the husband comes to her. All of the six troughs indicate pupil construction and occur during sad moments, which is the expected behavior of the pupils.

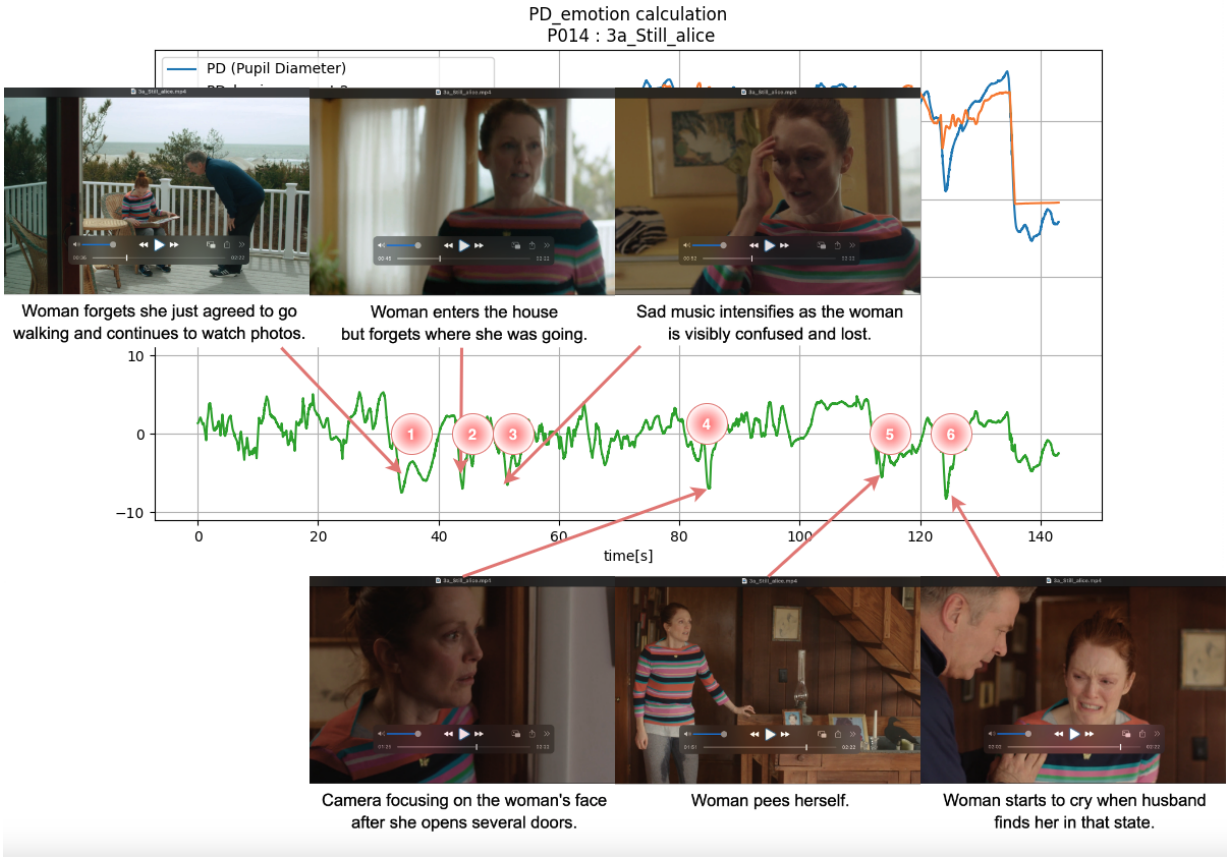


Figure 5.12: Analyzing the PD_{emotion} troughs and the corresponding scenes

5.2.5. Combining the pupil data of all participants

In figure 5.13, the PD_{emotion} component of their pupil diameter has been extracted from the total diameter values and plotted. It can be observed that, although the graph for each participant is different, they all follow the same general trend. An average of all the participants' data for the video "3a_Still_Alice" has been taken and plotted in figure 5.14. It can be observed that all the prominent peaks and troughs that are present in the combined (averaged) PD_{emotion} for all participants (figure 5.14) are also present in participant P014's data that have been analyzed in the previous figures 5.10 till 5.12 .

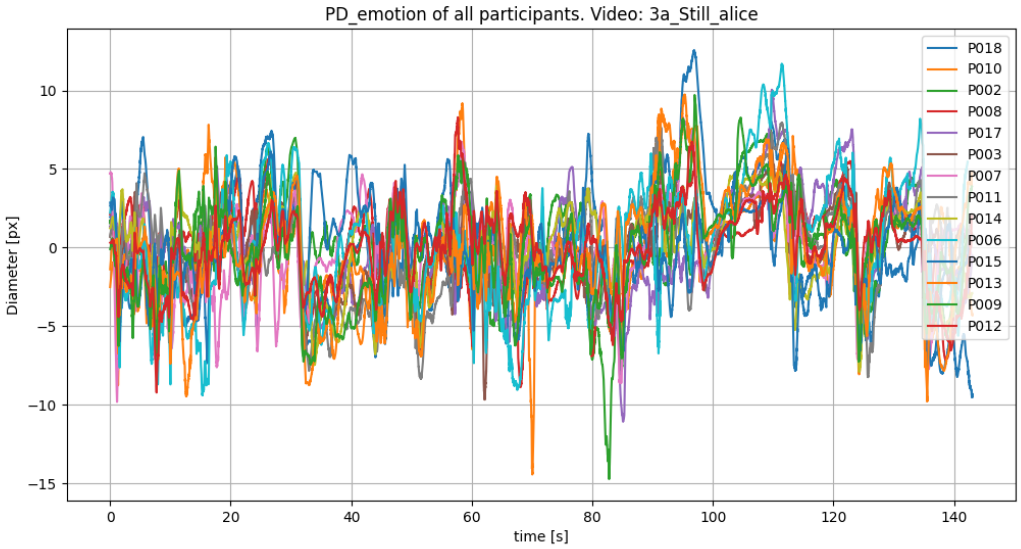


Figure 5.13: PD_{emotion} of all participants for video: “3a_Still_Alice”, emotion: Sadness

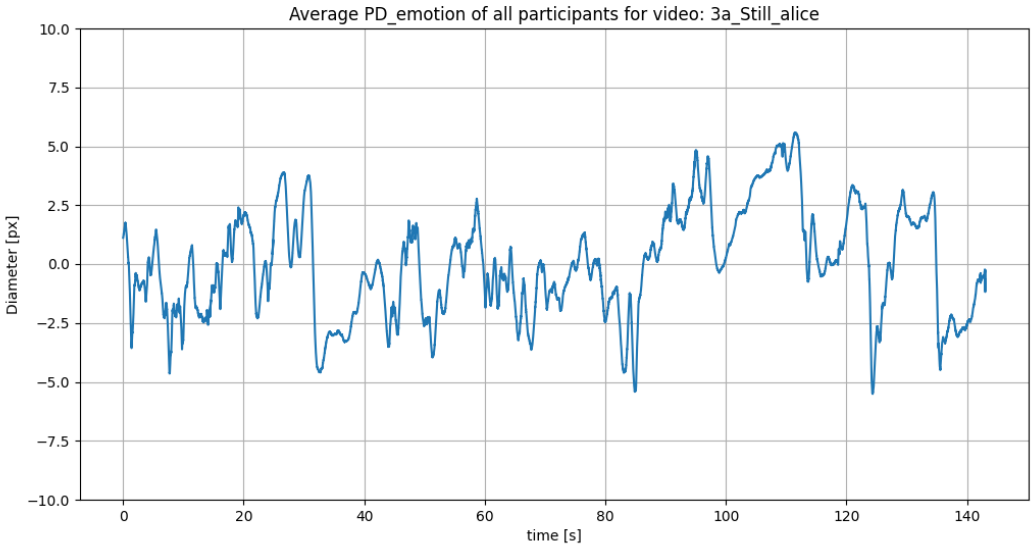


Figure 5.14: Mean of PD_{emotion} of all participants for video: “3a_Still_Alice”, emotion: Sadness

The same processing for the data done so far for the video “3a_Still_Alice”, was repeated for all the other videos. The PD_{emotion} component of the pupil data from all the participants and the videos were combined by averaging them and the distribution for each emotion has been plotted in figure 5.15. The box plots have been slightly shifted down so that the median diameter for the Neutral emotion falls on the x-axis for easier comparison between the emotions. It is evident from the previous analysis that labeling the data from the whole data as one emotion would be incorrect. So, for each video, the pupil data corresponding to only the specific segments assumed to evoke the basic emotions have been taken.

In the box plots in figure 5.15, it can be observed that the PD_{emotion} for the emotions Surprise, Happiness, Anger, Disgust and Fear are higher than for Neutral, and for Sadness, it is lower. This is in line with the results of a similar study done by Xue et. al, where it was found that *Low-Valance-Low-Arousal* videos

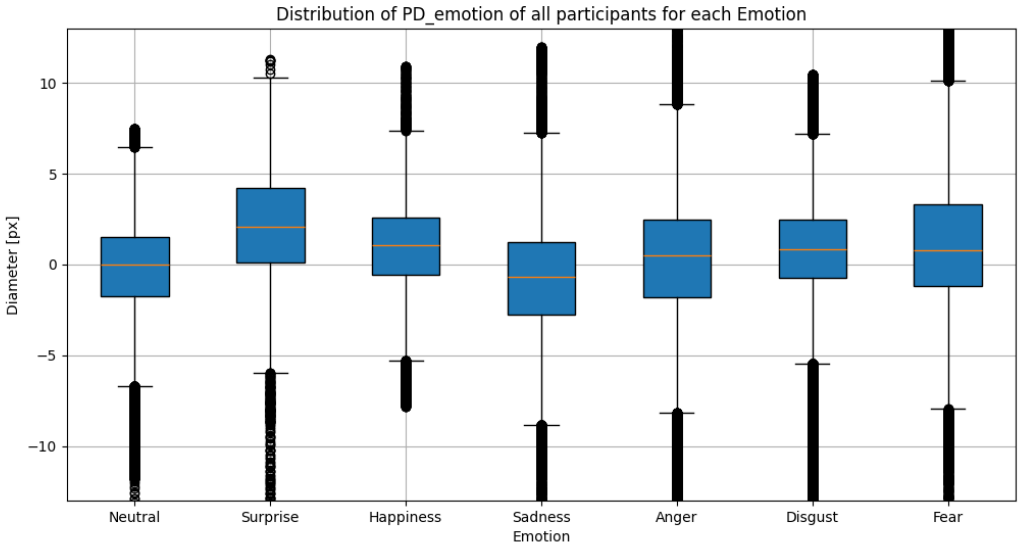


Figure 5.15: Box plots showing the distribution of the PD_{emotion} of all participants combined, for each emotion

cause the most pupil constriction and *High-Valance-High-Arousal* cause the most pupil dilation [17]. Figure 5.16, shows the translation of the basic emotions to the Valence-Arousal levels. Of the basic emotions, only Sadness falls in the category of Low-Valance-Low-Arousal, and figure 5.15 shows that the PD_{emotion} for Sadness is the lowest. For Surprise, the PD_{emotion} is the highest, which is also the emotion with the highest arousal level. For emotions Happiness, Anger, Disgust and Fear, the PD_{emotion} is higher than for Neutral but lower than for Surprise.

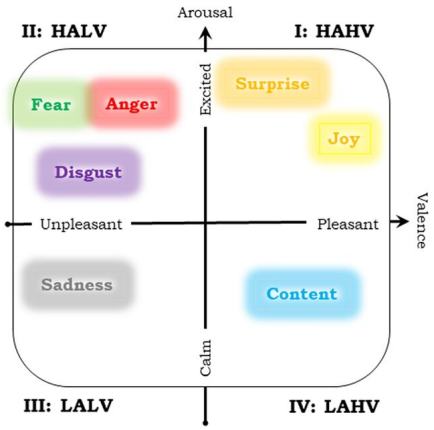


Figure 5.16: Basic emotions classified by Valence-Arousal levels [34]

5.2.6. Statistical Comparison tests

In order to check for significant differences in the pupil diameter across different emotions, statistical tests have been performed. Firstly, the Shapiro-Wilk and Lilliefors tests were performed to evaluate the normality of the pupil diameter data, which is important for deciding whether to use parametric or non-parametric tests for comparisons. The results of the normality tests are presented in Table 5.1. Both test indicate that the pupil diameter data are not normally distributed for all emotions ($p<0.05$). Although, the test statistics for the emotions Neutral, Happiness, Sadness and Anger suggest that their distribution is close to normal, as the Shapiro-Wilk test statistics are close to 1 and the Lilliefors test statistics are close to 0 for these

emotions. This could be because of the more sustained nature of these emotions over a period of time, as compared to Surprise, Fear and Disgust, which can have quick transient emotional responses.

	Shapiro-Wilk test		Lilliefors test	
Emotions	test stat	p-value	test stat	p-value
Neutral	0.9488	0.0000	0.0355	0.0010
Surprise	0.5982	0.0000	0.2051	0.0010
Happiness	0.9918	0.0000	0.0193	0.0010
Sadness	0.9962	0.0000	0.0200	0.0010
Anger	0.9924	0.0000	0.0153	0.0010
Disgust	0.6021	0.0000	0.2076	0.0010
Fear	0.6392	0.0000	0.1900	0.0010

Table 5.1: Normality test results for Shapiro-Wilk and Lilliefors tests

Since the pupil diameter data are not normally distributed, a non-parametric test has to be done to check for significant differences across the emotions. The Kruskal-Wallis H test has been chosen as it is suitable for comparing medians across more than two groups, and complements the visual comparison provided by the box plots in Figure \ref{fig:box_plot_diameter}, which illustrate differences in medians between the emotions. The result for the Kruskal-Wallis H test is given below. The large test stat value of 63596.165 and small p-value of 0.0 suggest that there are statistically significant differences between the medians of the emotion-related pupil diameter data.

$$Kruskal - Wallis\ H\ Test : stat = 63596.165, p = 0.0$$

Since the Kruskal-Wallis H test indicates only the existence of significant differences across the emotions but does not specify between which pairs of emotions these differences occur, a post-hoc analysis using Dunn's Test has been done. The results for Dunn's test using Bonferroni correction is shown in Table 5.2. The p-values which are less than 0.01 are colored green and indicate that a significant difference does exist between the emotions. The p-values which are more than 0.01 are colored red and indicate that a significant difference does not exist between the emotions. It can be observed that only the p-value for emotions Fear and Disgust are more than 0.01, meaning that no significant difference has been found between these two emotions. This could be because of the fact that participants tended to close their eyes or squint if they experienced intense fear, making the actual pupil size not being detected and in turn making the overall pupil size for fear smaller. An example of this can be seen in Figure 5.17, where the participant who reported a 9 for experienced emotional intensity kept their eyes closed for a few seconds several times during the video meant for evoking fear so the blink detection algorithm wasn't able to capture. Since the closing or squinting of the eyes is a natural part of the emotional reaction, the blink detection algorithm was not adjusted to reject this. In Figure 5.18, the pupil data for a participant who reported a 6 for fear can be seen. In this figure, the blink detection algorithm seems to be working well as the participant did not experience intense fear.

	Neutral	Surprise	Happiness	Sadness	Anger	Disgust	Fear
Neutral	-	0.0	0.0	0.0	0.0	0.0	0.0
Surprise	0.0	-	8.14e-192	0.0	0.0	0.0	0.0
Happiness	0.0	8.14e-192	-	0.0	0.0	2.16e-81	5.57e-77
Sadness	0.0	0.0	0.0	-	0.0	0.0	0.0
Anger	0.0	0.0	0.0	0.0	-	0.0	0.0
Disgust	0.0	0.0	2.16e-81	0.0	0.0	-	**1.0**
Fear	0.0	0.0	5.57e-77	0.0	0.0	**1.0**	-

Table 5.2: Pairwise comparison results for Dunn's Test

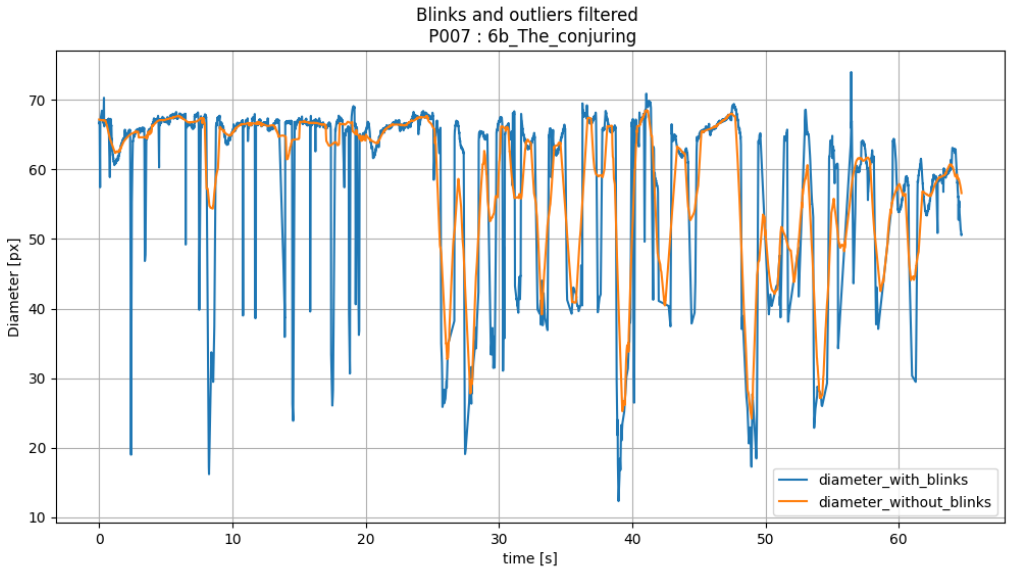


Figure 5.17: Participant P007's pupil diameter data for Fear stimuli. Before and After blink removal.

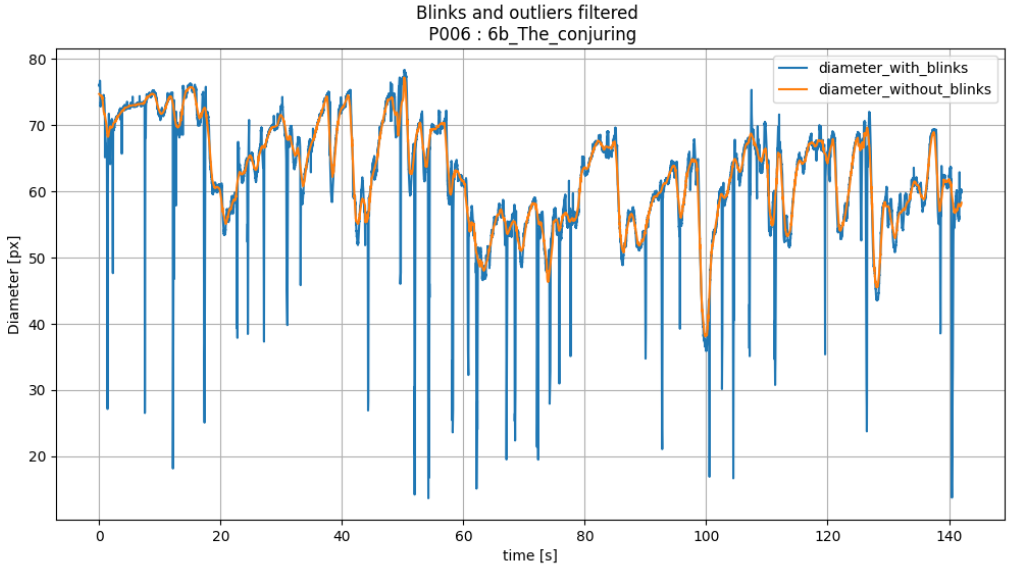


Figure 5.18: Participant P006's pupil diameter data for Fear stimuli. Before and After blink removal.

6

Conclusion

6.1. Summary

The project started with the goal of creating an emotional recognition dataset, which could be used for research and development of advanced emotion recognition systems using VR. Firstly, an analysis of the current situation in the field of emotion recognition using VR was done to identify a problem that could be addressed with this project. Then a possible solution to the problem was proposed and a clear objective for the project was set by looking at recent publications that align with the proposed solution and identifying what novel contributions could be made. Then the main components required to achieve the objective were identified and systematically implemented. These main components are the Stimuli for evoking emotions, the Data Collection System for recording the data containing emotional responses, and the Labeling Process for annotating the data.

The challenge encountered with the first component, the Stimuli, was the lack of a complete, publicly available and scientifically validated stimuli database designed to maximally utilize the emotion elicitation capabilities of VR. However, a good and effective alternative has been found with the selection of extensively studied video clips extracted from movies.

The challenge encountered with the second main component, the Data Collection System, was the lack of availability of a single platform designed to perform the complete data collection process efficiently. This has been addressed by integrating several open-source software tools and platforms into one system and automating the intermediate manual processes as much as possible to create a simple workflow. Thus making the process more efficient, consistent and less error-prone.

The challenge encountered with the third main component, the Labeling Process, was the complexity of accurately identifying all the segments that contain emotional responses, which are subjective and also mostly subconscious. The problem was that the participants could not be asked to continuously annotate the data as they engaged with the stimuli as this would be distracting and make the stimuli less effective. But, asking them after watching the video would pose the challenge of remembering all the scenes where they clearly felt an emotion. This has been addressed by providing the participants with some suggestions to help them recall their emotional responses and also giving them the option to add new segments or modify the current segments.

Also, validating the dataset was a challenge because of the subtle emotional responses captured in the data modalities using VR headset, which cannot be manually identified and would require the use of advanced emotion recognition systems that are not currently available. The solution to this has been found in the fact that the modalities captured have been shown to be correlated with emotions and the participants' self-reports show that the emotions have been elicitated effectively. Also, the pupil size which is one of the modalities that is a clear indicative of emotional arousal has been analyzed to confirm that the stimuli were effective in evoking the targeted emotions. A challenge in the pupil analysis was the dynamic effect of luminance from the VR display on the pupil diameter. This has also been addressed by developing a method to reduce the effect of luminance and isolate only the effect of emotions on pupil diameter.

6.2. Contributions

Through this project, two contributions have been made. First, the creation of a Data Collection System, which can be re-implemented and reused by other researchers. The Data Collection System was designed with reusability in mind by deciding to use commercially available and accessible equipment, and free open-source software and tools. The second contribution is the compiled dataset. This dataset can be used to conduct preliminary studies in the field of Emotion Recognition using VR and can be expanded using the data collection system.

6.3. Future Work

The future work consists of developing deep-learning models to accurately detect and identify the emotional responses captured in the different data modalities. The task however is not simple as it was observed that emotional responses can differ from person to person. Also, emotional responses are irregular and dynamic, changing rapidly over time. They are also expressed through a combination of multiple modalities which cannot be easily perceived manually. This makes the development of such models incredibly complex.

References

[1] Paul Ekman et al. “Nonverbal Leakage and Clues to Deception[†]”. In: *Psychiatry* 32.1 (Feb. 1969), pp. 88–106. DOI: 10.1080/00332747.1969.11023575.

[2] Björn Schuller et al. “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture”. In: vol. 1. June 2004, pp. I–577. DOI: 10.1109/ICASSP.2004.1326051.

[3] Andrea Kleinsmith et al. “Affective Body Expression Perception and Recognition: A Survey”. In: *IEEE Transactions on Affective Computing* 4.1 (2013), pp. 15–33. DOI: 10.1109/T-AFFC.2012.16.

[4] R.W. Picard et al. “Toward machine emotional intelligence: analysis of affective physiological state”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.10 (2001), pp. 1175–1191. DOI: 10.1109/34.954607.

[5] J.A. Healey et al. “Detecting stress during real-world driving tasks using physiological sensors”. In: *IEEE Transactions on Intelligent Transportation Systems* 6.2 (2005), pp. 156–166. DOI: 10.1109/TITS.2005.848368.

[6] Soujanya Poria et al. “A review of affective computing: From unimodal analysis to multimodal fusion”. In: *Information Fusion* 37 (2017), pp. 98–125. DOI: https://doi.org/10.1016/j.inffus.2017.02.003. URL: https://www.sciencedirect.com/science/article/pii/S1566253517300738.

[7] D. Freeman et al. “Virtual reality in the assessment, understanding, and treatment of mental health disorders”. In: *Psychological Medicine* 47.14 (Mar. 2017), pp. 2393–2400. DOI: 10.1017/s003329171700040x.

[8] G. McKeown et al. “The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent”. English. In: *IEEE Transactions on Affective Computing* 3.1 (Jan. 2012), pp. 5–17. DOI: 10.1109/T-AFFC.2011.20.

[9] Raquel Ros et al. “Child-robot interaction in the wild: Advice to the aspiring experimenter”. In: Nov. 2011, pp. 335–342. DOI: 10.1145/2070481.2070545.

[10] Javier Marín-Morales et al. “Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing”. In: *Sensors* 20.18 (2020). URL: https://www.mdpi.com/1424-8220/20/18/5163.

[11] Paul Ekman et al. *Facial action coding system: A technique for the measurement of facial movement*. 1st. Academic Press, 1978.

[12] Timo Partala et al. “Pupil size variation as an indication of affective processing”. In: *International Journal of Human-Computer Studies* 59.1 (2003). Applications of Affective Computing in Human-Computer Interaction, pp. 185–198. DOI: https://doi.org/10.1016/S1071-5819(03)00017-X. URL: https://www.sciencedirect.com/science/article/pii/S107158190300017X.

[13] Atanu Samanta et al. “On the role of head motion in affective expression”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 2886–2890. DOI: 10.1109/ICASSP.2017.7952684.

[14] Anton Batliner et al. “The Automatic Recognition of Emotions in Speech”. In: Jan. 2011, pp. 71–99. DOI: 10.1007/978-3-642-15184-2_6.

[15] Luma Tabbaa et al. “VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5.4 (Dec. 2022). DOI: 10.1145/3495002. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3495002.

[16] Yingying Zhao et al. “Do Smart Glasses Dream of Sentimental Visions?: Deep Emotionship Analysis for Eyewear Devices”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1 (Mar. 2022), pp. 1–29. DOI: 10.1145/3517250. URL: http://dx.doi.org/10.1145/3517250.

[17] Tong Xue et al. “CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° VR Videos”. In: *IEEE Transactions on Multimedia* 25 (2023), pp. 243–255. DOI: 10.1109/TMM.2021.3124080.

[18] SREEJA P S et al. “Emotion Models: A Review”. In: *International Journal of Control Theory and Applications* 10 (Jan. 2017), pp. 651–657.

[19] Anvarjon Tursunov et al. “Discriminating Emotions in the Valence Dimension from Speech Using Timbre Features”. In: *Applied Sciences* 9.12 (2019). DOI: 10.3390/app9122470. URL: https://www.mdpi.com/2076-3417/9/12/2470.

[20] Javier Marín-Morales et al. “Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors”. In: *Scientific Reports* 8 (Sept. 2018). DOI: 10.1038/s41598-018-32063-4.

[21] Joris Heyse et al. “A Personalised Emotion-Based Model for Relaxation in Virtual Reality”. In: *Applied Sciences* 10.17 (2020). DOI: 10.3390/app10176124. URL: https://www.mdpi.com/2076-3417/10/17/6124.

[22] Christos Kyrlitsias et al. “Social Conformity in Immersive Virtual Environments: The Impact of Agents’ Gaze Behavior”. In: *Frontiers in Psychology* 11 (2020). DOI: 10.3389/fpsyg.2020.02254. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.02254.

[23] Benjamin J. Li et al. “A Public Database of Immersive VR Videos with Corresponding Ratings of Arousal, Valence, and Correlations between Head Movements and Self Report Measures”. In: *Frontiers in Psychology* 8 (2017). DOI: 10.3389/fpsyg.2017.02116. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.02116.

[24] Pierre Philippot. “Inducing and assessing differentiated emotion-feeling states in the laboratory”. In: *Cognition and Emotion* 7.2 (1993). PMID: 27102736, pp. 171–193. DOI: 10.1080/02699939308409183. eprint: https://doi.org/10.1080/02699939308409183. URL: https://doi.org/10.1080/02699939308409183.

[25] Barbra Zupan et al. “Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion”. In: *The Journal of Social Psychology* 160.6 (2020). PMID: 32419668, pp. 768–789. DOI: 10.1080/00224545.2020.1758016. eprint: https://doi.org/10.1080/00224545.2020.1758016. URL: https://doi.org/10.1080/00224545.2020.1758016.

[26] Gyouhyung Kyung et al. “Curved Versus Flat Monitors: Interactive Effects of Display Curvature Radius and Display Size on Visual Search Performance and Visual Fatigue”. In: *Human Factors* 63.7 (2021). PMID: 32374635, pp. 1182–1195. DOI: 10.1177/0018720820922717. eprint: https://doi.org/10.1177/0018720820922717. URL: https://doi.org/10.1177/0018720820922717.

[27] Pin-Sung Ku et al. “PeriText: Utilizing Peripheral Vision for Reading Text on Augmented Reality Smart Glasses”. In: Mar. 2019, pp. 630–635. DOI: 10.1109/VR.2019.8798065.

[28] Moritz Kassner et al. “Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction”. In: *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Apr. 2014). DOI: 10.1145/2638728.2641695.

[29] Ali Mollahosseini et al. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild”. In: *IEEE Transactions on Affective Computing* 10.1 (Jan. 2019), pp. 18–31. DOI: 10.1109/taffc.2017.2740923.

[30] Yih-Giun Cherng et al. “Background luminance effects on pupil size associated with emotion and saccade preparation”. In: *Scientific Reports* 10.1 (2020), p. 15718. DOI: 10.1038/s41598-020-72954-z. URL: https://doi.org/10.1038/s41598-020-72954-z.

[31] Paweł Tarnowski et al. “Eye-Tracking Analysis for Emotion Recognition”. In: 2020 (Jan. 2020). DOI: 10.1155/2020/2909267. URL: https://doi.org/10.1155/2020/2909267.

[32] Sebastiaan Mathôt. “Pupillometry: Psychology, Physiology, and Function”. In: *Journal of Cognition* 1 (2018). URL: https://api.semanticscholar.org/CorpusID:54068639.

[33] Jing Zhang et al. “Significant Measures of Gaze and Pupil Movement for Evaluating Empathy between Viewers and Digital Content”. In: *Sensors* 22.5 (2022). DOI: 10.3390/s22051700. URL: https://www.mdpi.com/1424-8220/22/5/1700.

[34] Majid Riaz et al. “High dynamic range multimedia: better affective agent for human emotional experience”. In: *Multimedia Tools and Applications* (Aug. 2023), pp. 1–16. DOI: 10.1007/s11042-023-16524-1.