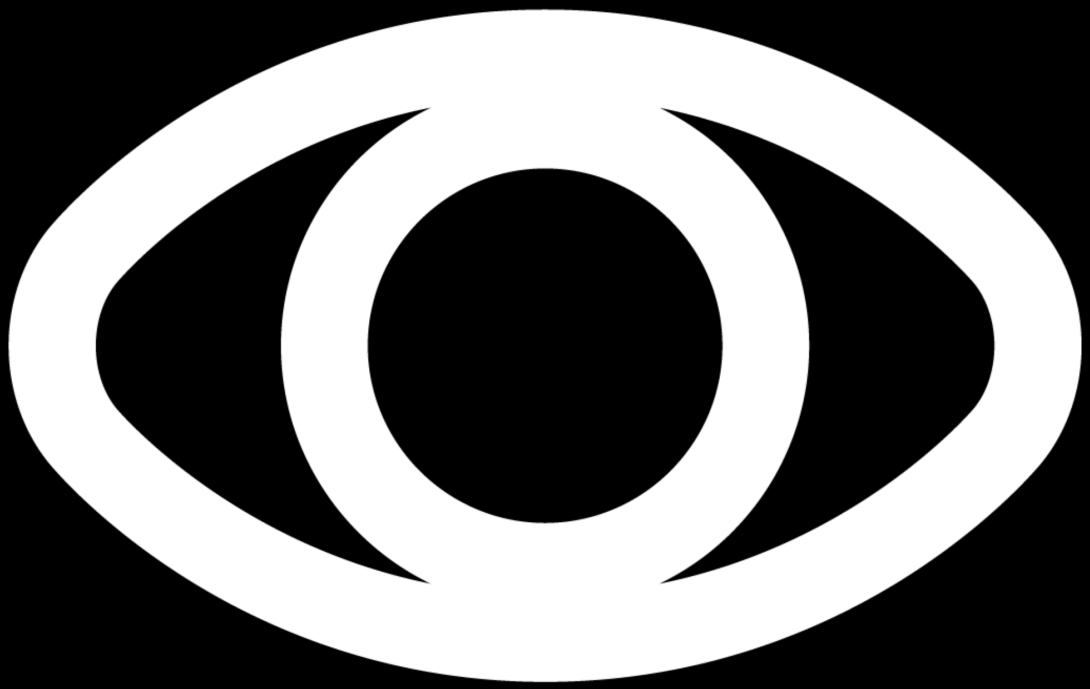
Multi-representation Emotion Recognition in Immersive Environments

MSc Thesis

Tongyun Yang

Delft University of Technology



TU Delft

Multi-representation Emotion Recognition in Immersive Environments

by

Tongyun Yang

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday October 22, 2024 at 15:00 PM.

Student number:	5651794	
Project duration:	Jan., 2024 – Oct., 2024	
Supervisor:	Dr. Guohao Lan,	TU Delft, Embedded Systems
Thesis committee:	Prof.dr. K.G. (Koen) Langendoen,	TU Delft, Embedded Systems
	Dr. Guohao Lan,	TU Delft, Embedded Systems
	Dr. Xucong Zhang,	TU Delft, Computer Vision Lab

Cover: Eye image downloaded from pngfind.com, and cover made by the author Tongyun Yang

TU Delft

Preface

Writing this preface marks the end of my Master's journey, a cloth woven with experiences and memories that I will cherish for a lifetime. During a farewell with friends, I described them as shooting stars streaking across my sky one night. I may not know where they came from or where they are heading to, but that matters little. What matters is the brilliance they cast upon that night, and that is enough. I cannot recall whether I have had nights like this before, but I am certain that more such nights await me in the future. That night, the sky adorned with lights, is something I will treasure forever.

I want to express my gratitude to ***Guohao*** *for all he has given to me during this time, from guidance on my thesis to mindset towards life. He often describes himself as a strict person, and indeed he is, he tolerates nothing less than perfection. Yet, he is also remarkably gentle, gentle in a way that I could always perceive without the need for words.*

I also want to thank ***Xucong*** *for his support in every context. He is direct, he is optimistic, and he has the ability to cheer up the spirits of those surrounding him. If I one day find myself in a position like his today, he will forever remind me of how to approach the world around me.*

Lastly, my deepest thanks are for my mother. Though it has been years since I last saw her in person, each time we call she soothes my restless heart. Her words of wisdom have always been my compass while I sail towards the unknown future. She is the love of my life, now and always.

Tongyun Yang
Delft, October 2024

Contents

Abstract

This study addresses the gap for fine-grained emotion recognition in immersive environments utilizing solely data from on-board sensors. Two data representations of users eyes are utilized, including periocular recordings and eye movements (gaze estimation and pupil measurements). A novel multi-representation method integrating feature extractors for each representation alongside an effective feature fusion technique is proposed. The method significantly outperforms baselines that use only a single representation or incorporate content stimuli. It achieves an F1-score of 0.85 with 10% data, approximately 40 seconds of data from all emotions, for personal adaptation, recognizing emotions while watching unseen parts of the stimuli used for adaptation. In a more practical scenario, the method achieves an F1-score of 0.71 with five seconds of personal adaptation data from each emotion, recognizing emotions while watching completely unseen stimuli. Under the same but more extreme condition, where only one second of data is available, the proposed achieves an F1-score of 0.68. Furthermore, the study demonstrates that estimated labels can substitute for user-provided labels without sacrificing performance in emotion recognition, thus eliminating the need for users to manually label emotion elicitation segments. Future work will focus on improving performance by allocating more computational resources and making architectural modifications, conducting deeper investigations into the decision-making process, and developing real-time recognition systems for in-the-wild experiments. The results of this study suggest that more engaging, adaptive, and personalized experiences in immersive environments can be developed.

1 Introduction

1.1. Background & Motivation

Emotions are part of human nature, and a significant portion of the human brain is dedicated to understanding and processing them. They occupy daily life to a greater extent than realized. During social interactions, the human brain captures and interprets various cues such as facial expressions, body movements, and speech tone, trying to understand each other's emotional states. During movies, filmmakers strategically evoke emotional responses in audiences, often eliciting laughter to induce happiness, and in contrast, provoking tears to trigger sympathy. The inability to understand these emotional nuances is often associated with social deficits, a symptom in diagnosing conditions such as autism spectrum disorder [63]. Understanding in this context means more than grasping fundamentals of some knowledge, it means having empathy with other's emotions.

The importance of understanding emotions extends beyond social interaction. Emotions play a critical role in the evolution of consciousness and the operations of all mental processes [34]. Understanding emotions could potentially provide insights into unresolved questions in psychology and neuroscience, particularly in areas such as decision-making, problem-solving, and other cognitive functions. Moreover, emotional understanding could lead to better child education, improvements in social interaction and effective mental health treatments.

The recent emerging interest in immersive environments has led to a growing amount of research on emotion understanding within these contexts [30, 85, 96, 100, 101]. Immersive environments has the advantage of offering experiences that closely mimic the physical world while allowing for isolated and controlled experimental conditions. This makes immersive environments, such as virtual reality, particularly valuable for emotion research. The ability to systematically control environmental factors and stimuli while maintaining the same ecological responses as the physical world, provides more potential than in traditional laboratory settings. However, the field remains largely unexplored. Table 1.1 presents the comparison of recent studies across several key characteristics in emotion understanding, more specifically, emotion recognition. The key characteristics are important as they maximize the validity of the elicited emotions and minimize the effect of unrelated factors, leading to a more comprehensive emotion state of subjects. Hence, effective emotion recognition should ideally be conducted in immersive environments, employing effective emotion elicitation methods, utilizing only on-board sensors, and aiming to recognize fine-grained emotions. These criteria are established based on the need for controlled environments that can successfully trigger emotions without compromising the natural setup of equipped devices, typically consisting of a single headset. Additionally, the identification of fine-grained emotions is crucial for a nuanced understanding. The comparison shows that current studies does not fully satisfy the defined characteristics, thus highlighting the need for the present study to address this gap.

This study builds upon the work of Bishwas [74], who developed a system for data collection in immersive environments using only on-board sensors to capture fine-grained emotions elicited by effective stimuli. The current study initially focuses on processing the collected raw data to mitigate noise, aiming to extract more meaningful information and constructing datasets suitable for efficient model training. Subsequently, the study proposes a model that incorporates various data types as input for

Research	Immersive Env.	Fine-grained Emo.	Effec. Stimuli	On-board Sen.
RCEA [101]	×	×	✓	✓
RCEA-360 [96]	✓	×	✓	✓
VREED [85]	✓	×	✓	×
Total VREcall [30]	✓	×	✓	×
Blink of an Eye [100]	×	✓	×	×
SEED-V [51]	×	✓	✓	×
DECAF [1]	×	×	✓	×
DEAP [41]	×	×	✓	×
Excitement Detect [2]	✓	×	✓	✓
Arousal Detect [91]	✓	×	✓	×

Table 1.1: Comparison of different studies in emotion recognition, the characteristics are shortened from Immersive Environments, Fine-grained Emotions, Effective Stimuili and On-board Sensors.

fine-grained emotion recognition. The proposed model is trained and evaluated using the constructed datasets.

1.2. Challenges & Research Objectives

The challenges in this study primarily relate to two aspects. The first involves effectively utilizing diverse data types for emotion recognition, and the second concerns developing methods for accurate emotion recognition with limited data.

The first challenge arises from the complexities of feature fusion. Though progress has been made in achieving progress in performance through feature fusion, it remains an open area that requires further research [50]. Different fusion techniques deliver varying results, developing and employing the optimal method is a challenging task. The second challenge stems from the practical application of emotion recognition in real-world scenarios. While having access to users content, such as viewing videos or reading texts, could facilitate emotion recognition, it raises privacy concerns. It also estimates emotions associated with specific content rather than recognizing emotions directly from users responses. Hence, the recognition process should solely rely on users' behavioral responses. Moreover, the developed method should require a minimal amount of data from each user for adaptation, thereby enhancing the user experience. In order to address these challenges, the research objectives of this study are as follows:

1. To create and validate a comprehensive dataset of emotional responses in immersive environments, incorporating different types of collected data.
2. To develop and evaluate a method for emotion recognition that effectively utilizes diverse data types collected in immersive environments.
3. To design and implement a fusion technique that optimally integrates multiple data types for improved emotion recognition accuracy.
4. To compare the performance of the developed privacy-preserving method against approach that utilizes content information.
5. To investigate and evaluate an approach that achieves robust emotion recognition performance with limited data used for personal adaption.

1.3. Contributions

The main contribution of this study is that it is the first study that delivers promising fine-grained emotion recognition performance utilizing solely on-board sensors data gathered in immersive environment triggered by effective stimuli. First, a comprehensive dataset is constructed in immersive environments, collecting data from 20 subjects by eliciting each with effective stimuli across seven emotions. The data is gathered exclusively from on-board sensors of the device equipped on the headset. Then, a model incorporating multiple data types is proposed and trained on the constructed dataset, achieving promising performance. The contribution fills in the gap in recently studies conducted in emotion recognition, as it satisfies all the characteristics defined in Table 1.1. The contribution in this study

along the previous work conducted by Bishwas [74], is combined to form an academic paper aiming for submission to a top-tier conference.

1.4. Thesis Overview

In Chapter 2, a comprehensive review of related work is first presented, covering emotion models, immersive environments and emotions, existing emotion datasets, and state-of-the-art emotion recognition methods. Following this, a summary of the data collection process is presented in Chapter 3, including the emotion model used, the dataset collection tool, and the characteristics of the collected raw data. After that, Chapter 4 presents comprehensive details of the data processing and dataset preparation steps are presented, addressing challenges in raw data and outlining the methods used to create the final dataset structure. The proposed emotion recognition method is then presented in Chapter 5. Details of the overall architecture and different modules extracting features from different types of data are presented, along with the implemented effective feature fusion technique. The process of data flowing from input to the recognized emotion is comprehensively described. Subsequently in Chapter 6, the experimental design, implementation details, and performance evaluation results are outlined. This chapter also includes ablation studies and profiling of the proposed approach. A discussion of the results follows in Chapter 7, providing more insights to the decision-making process and addressing the study's limitations. The study concludes in Chapter 8 by summarizing the key findings and contributions. It also suggests directions for future work and discusses the implications of this research for human-computer interaction in immersive environments.

2

Related Work

2.1. Emotion Models

Emotion models are broadly categorized into two types: categorical and dimensional [67]. Categorical models require the selection of a single emotion from a predefined set, representing the most appropriate feeling conveyed, such as Ekman and Friesen's six basic emotions [24] and Izard's ten core emotions [35]. Dimensional models, in contrast, employ quantitative measures through multidimensional scaling. Each dimension represents a specific feature of human emotion, and the combination provides an interpretation of the emotional state. The Circumplex Model of Affect by Russell [76] utilizes two dimensions: valence and arousal. In order to differentiate closely related emotions, the pleasure-arousal-dominance (PAD) model introduces dominance as a third dimension [60]. To quantify these dimensional scales, tools such as self-assessment manikin (SAM) [10] and Feeltrace [15] are employed.

Recent studies that involve defining emotional models usually employ a combination of both categorical and dimensional models. For example, Buechel and Hahn [11] utilized a model combining the categorical approach of Ekman's six basic emotions with the dimensional valence-arousal model for emotion analysis in text. Similarly, Tzirakis et al. [87] used a hybrid approach integrating categorical emotions and dimensional features for multi-modal emotion recognition in videos. These studies demonstrate the effectiveness of leveraging both categorical specificity and dimensional variability in emotion recognition applications.

2.2. Immersive Environments and Emotions

In recent years, immersive environments have gained significant attention in both academic research and industrial applications [70, 88]. Immersive environments include Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). They offer users experiences that blur the boundary between the digital and physical worlds [26]. The significantly growing interest in immersive technologies is driven by the increasing need to enhanced remote collaboration tools. Especially during the trend towards distributed work and study after the recent global pandemic, namely Covid-19 [53, 59]. The key characteristic of immersive environments is the ability to create a sense of presence, allowing users to feel as if they are physically present in the virtual space [70].

Research has shown that immersive environments can elicit a wide range of emotions in users [70], leading to applications across various fields such as psychology, education, and entertainment. In psychology, immersive environments have been used for exposure therapy to treat phobias and anxiety disorders [16, 20]. Educational institutions have employed these technologies to provide immersive learning experiences for enhancing knowledge retention and understanding [77, 83]. The entertainment industry has also leveraged immersive environments to create more emotionally engaging storytelling experiences [73].

Several studies have revealed a strong relationship between immersive environments and emotional experiences. Pavic et al. [70] found that the level of immersion directly correlates with the intensity of emotional responses, while Diemer et al. [19] showed that virtual environments can induce physiological responses similar to those experienced in real-world emotional situations. However, it is

important to note that the emotional impacts in immersive environments are not always positive. Some studies have raised concerns about the potential harm caused by these technologies, such as emotional distress or anxiety, particularly in highly realistic and intense scenarios [46]. Therefore, thoughtful design and careful consideration of the emotional impact on users should be taken into account when developing immersive experiences.

2.3. Emotion Datasets

The development and evaluation of emotion recognition heavily depend on comprehensive datasets. Various emotion datasets consisting of different modalities and utilizing different collection methods are collected by researchers serving the same purpose of recognising different emotions.

One of the most utilized modalities in emotion recognition is numerical time series data representing users' behavioral and physiological signals. Several datasets have been collected in immersive environments using similar methods. For instance, VREED [85] includes eye movements (gaze estimation and pupil diameter measurements), electrocardiography (ECG), galvanic skin response (GSR), and self-report data from 34 participants viewing 12 videos in virtual reality (VR). Similarly, PEM360 [29] includes identical data types as VREED, with the addition of head movements and heart rate, collected from 32 participants watching eight videos. CEAP-360VR [95] consists of the same data types as PEM360 from 32 participants viewing eight videos in VR, with reported motion sickness and presence levels in addition. SEED-IV [104] gathered electroencephalography (EEG) and eye movements from 15 subjects while being exposed to six clips. These datasets typically involve participants being equipped with head-mounted displays (HMDs) and additional devices for capturing physiological signals. They present participants with stimuli aiming to trigger certain emotions in the immersive environment, and the collected data are labeled based on users' self-reported emotional responses.

Facial expressions provide cues for emotion recognition as well, leading to the development of several image-based datasets. FER+ [6] comprises 28,709 facial expression images sourced from the internet with multi-label annotations across seven emotions. AffectNet [61] includes 440,000 images from the internet and manually annotated with single-label emotions across seven emotions which are identical as FER+. 4DFAB [14] was collected over five years and consists of 1.8 million 3D facial meshes from 180 subjects. The expressions in 4DFAB were captured through both posed emotions and spontaneous reactions to video stimuli. Similar to image, video provides visually informative content as well. However, it has an additional temporal axis, offering potential advantages over images for emotion recognition. MAFW [52] consists of multi-modal clips from diverse sources such as movies, TV dramas, and short videos. MAFW features subjects expressing a wide range of emotions in various scenarios, providing contextual information for emotion recognition.

The emotion datasets discussed can be broadly categorized into two types based on their collection methodologies: elicited emotion datasets (e.g., VREED, PEM360, CEAP-360VR, SEED-IV) and enacted emotion datasets (e.g., FER+, AffectNet, 4DFAB, MAFW). Elicited emotion datasets are collected in controlled laboratory settings where participants are exposed to designed stimuli to trigger genuine emotional responses under standardized conditions. In contrast, enacted emotion datasets are sourced from the internet or involved participants deliberately posing expressions of certain emotions, which may not reflect genuinely experienced affective states. Elicited emotion datasets often reveal more nuanced relationships between emotions and biological responses, potentially capturing subtle physiological and behavioral changes associated with different emotions. Enacted emotion datasets, while possibly including a mix of genuine and posed expressions, typically capture more obvious differences in human emotional expressions across a wider range of contexts.

2.4. Emotion Recognition Methods

The significant performance gain brought by deep learning have revolutionized various domains of artificial intelligence, achieving superior results in natural language processing (NLP) [18, 89], computer vision (CV) [21, 31, 43], and other fields. This paradigm shift has influenced emotion recognition research, as several modalities involved align with those benefiting from deep learning advancements.

Traditionally, emotion recognition methods have focused on uni-modal approaches, primarily utilizing one modality among facial expressions, speech audio, and physiological signals. In facial expression-based methods, Siqueira et al. [81] demonstrated an accuracy of 87.15% on the FER+ dataset (discussed in Section 2.3) with convolutional neural networks (CNNs). For speech-based emotion recog-

nition, recurrent neural networks (RNNs) and their variants, particularly long short-term memory (LSTM) networks, have shown promising results in capturing temporal dependencies. Zhao et al. [103] achieved over 95% accuracy across seven emotions on the Berlin EmoDB dataset [12] on both speaker-dependent and speaker-independent scenarios. Methods utilizing physiological signals have also benefited from deep learning advancements. Zhong et al. [105] achieved an accuracy of 73.84% in classifying four emotions using graph neural networks (GNNs) on the SEED-IV dataset (mentioned in Section 2.3). They effectively modeled the spatial relationships between different brain regions by leveraging the inherent graph-like structure of EEG electrode placements.

While uni-modal approaches have made great progress in emotion recognition, there is a growing trend towards multi-modal emotion recognition methods. These methods align with human emotion recognition processes during personal interactions. They leverage the complementary information from different modalities to achieve more robust and accurate emotion recognition. For instance, Makhmudov et al. [56] combined speech audio with corresponding transcribed text, achieving an accuracy of 67.81% on the MELD dataset [71]. Tan et al. [86] demonstrated a recognition rate of 83.33% by utilizing facial expressions and EEG. Ma et al. [54] achieved an accuracy of 73.95% by integrating speech audio, transcribed text, and facial expression video, which is currently the highest accuracy achieved on the IEMOCAP dataset [13].

However, it is important to note that multi-modal methods do not always yield superior results. For example, in studies on the MELD dataset [71], while Yun et al. [97] achieved 67.37% accuracy utilizing speech audio, transcribed text, and facial expression video, Xue et al. [94] achieved state-of-the-art performance on the same dataset using only the transcribed text. This highlights the challenge of effectively fusing features from different modalities. To address a similar issue, Zhang et al. [102] proposed a novel attention-based fusion mechanism that dynamically weights the contributions of different modalities based on their relevance to the emotion recognition task.

As emotion recognition systems become potentially prevalent in real-world applications, the need for real-time processing has gained attention [17]. Studies have focused on developing lightweight models and efficient inference techniques to enable emotion recognition on edge devices with limited computational resources [57, 68, 69]. This trend reflects the growing demand for practical and deployable emotion recognition solutions.

2.5. Numerical Time Series Classification

Emotion recognition in immersive environments utilizing multi-modal data involves analyzing time series data, such as eye movements, ECG, and heart rate collected from sensors to recognize emotional states. This task falls under the broader category of numerical time series classification, which has been extensively studied [5, 33].

Traditional methods for time series classification include distance-based methods such as Dynamic Time Warping (DTW) [7], feature-based methods like extracting statistical features [64], and ensemble methods such as BOSS [78]. These traditional methods have been successfully applied to various time series classification problems and serve as strong baselines. However, they often require hand-crafted features and may struggle to capture complex temporal patterns in the data.

In recent years, there has been a growing trend towards using deep learning techniques for time series classification [33]. RNNs have shown promising results in capturing temporal dependencies and learning discriminative features from raw time series data [37, 66]. Transformer-based methods have also gained attention, with Zerveas et al. [99] proposing a framework for multivariate time series representation learning that achieves state-of-the-art performance on multiple datasets.

Empirical studies comparing the performance of traditional methods and deep learning approaches for time series classification have found that deep learning models generally outperform traditional methods, especially on complex and large-scale datasets [33]. However, the performance gain varies depending on the specific dataset and problem domain. Deep learning models have the advantage of automatically learning relevant features from raw data, eliminating the need for manual feature engineering. They can also capture intricate patterns and long-range dependencies that traditional methods may struggle with.

In the context of emotion recognition from multi-modal numerical time series data, state-of-the-art methods often employ deep learning architectures. For example, Ma et al. [55] proposed a recurrent network with multi-modal data as input, such as EEG and ECG, using an LSTM network equipped

with residual connections for emotion classification. Li et al. [49] developed a multi-modal emotion recognition system that integrates EEG and other physiological signals using a hierarchical attention-based CNN network to capture both spatial and temporal dependencies in the data.

2.6. Video Understanding

Over the past decade, significant progress has been achieved in the field of video understanding. Early approaches relied on hand-crafted features and traditional machine learning algorithms, where the focus was on extracting local space-time features from videos and using them for action recognition. For example, Laptev et al. [44] proposed an approach that detects space-time interest points in videos and described them using local spatial and temporal features, which were then used to train a support vector machine (SVM) classifier for action recognition. However, with the advent of deep learning, researchers have developed methods capable of learning hierarchical representations directly from raw video data. Karpathy et al. [38] introduced a deep learning framework for large-scale video classification that learns spatio-temporal features using CNNs. Their approach involves extending the CNN kernel in time domain and training it on a large dataset of labeled videos, enabling the network to learn discriminative features for video across different context.

Video understanding contributes to a wide range of tasks, including action recognition, event detection, video captioning, etc. Action recognition aims to classify human actions in videos, such as walking, running, or jumping [80]. Event detection focuses on identifying specific events or activities, like a birthday party or a football match [93]. Video captioning generates natural language descriptions of video content, bridging the gap between visual and textual information [90].

In recent years, video understanding has been applied to various domains beyond traditional computer vision tasks. For instance, it has been used in healthcare for analyzing medical videos and assisting with diagnosis [25]. In the field of robotics, video understanding enables robots to perceive and interact with their environment more effectively [47].

Notably, video has emerged as a valuable modality for emotion recognition. Several works have explored the use of video data to predict emotional states. Kahou et al. [36] proposed a multi-modal approach that combines facial expressions, audio, and textual information from videos to classify emotions with CNNs. Ebrahimi et al. [22] extended the work by incorporating spatio-temporal features extracted from video frames using a hybrid RNN-CNN architecture.

The potential of video understanding in immersive environments for emotion recognition is an emerging area of research [62]. Integrating video data with other modalities, such as physiological signals and user self-reports, could lead to more accurate and robust emotion recognition systems in immersive settings.

3

Emotion Dataset Collection

3.1. Emotion Model

As discussed in Section 2.1, emotion models are typically classified as either categorical or dimensional. Categorical models, such as Ekman and Friesen's six basic emotions [24], require the selection of a single emotion from a predefined set. While dimensional models, like the Circumplex Model of Affect by Russell [76], employ quantitative measures through multidimensional scaling. This study integrates elements from both categorical and dimensional models. Ekman and Friesen's six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) are adopted along with 'neutral' as the foundation of the emotion representations. The inclusion of the 'neutral' state allows for a more comprehensive capture of the emotional states experienced by users, particularly in situations where none of the six basic emotions are strongly present. Additionally, an intensity scaling dimension is included to capture the self-reported intensity of each emotion experienced by the users. It is important to note that the intensity scaling dimension is only utilized for self-reported user emotions and does not play a role in the emotion recognizing stage. Figure 3.1 illustrates the adapted emotion categorical representation utilized.

3.2. Dataset Collection Tool

The emotion collection tool used in this research was developed by Bishwas Regmi as part of his Master's thesis [74]. The tool is designed to collect multi-modal data from participants while they are presenting with emotion-eliciting video stimuli in a VR environment. The experimental setup involves presenting a total of 14 video stimuli to each participant, with pairs of stimuli (labeled as (a) and (b)) intending to evoke the same emotion. This design results in data collection for seven distinct emotions, each comprising two sessions (a and b) corresponding to the paired stimuli.

The collected data includes eye movements (gaze estimation and pupil diameter measurements), periocular recordings from both eyes, head movements, and speech audio. Additionally, the tool incorporates a self-report mechanism, allowing participants to provide subjective emotion intensity ratings for various segments of each presented stimulus. These ratings are collected on a scale ranging from 1 to 10, where 1 indicates no intensity, 6 represents mild intensity, and 10 signifies very high intensity.

This data collection approach enables the collection of a dataset that captures both objective behaviour responses and subjective emotional experiences, providing a solid foundation for emotion recognition in the further stages.

3.2.1. Hardware Components

The emotion collection tool utilizes the following hardware components:

VIVE Pro VR headset: Creates a standardized VR environment for presenting the video stimuli to the participants.

Pupil Labs eye-tracking add-on: Integrated with the VIVE Pro headset to capture eye movement data and record the periocular region of both eyes [39].

Integrated microphone: Records speech audio data from the participants.

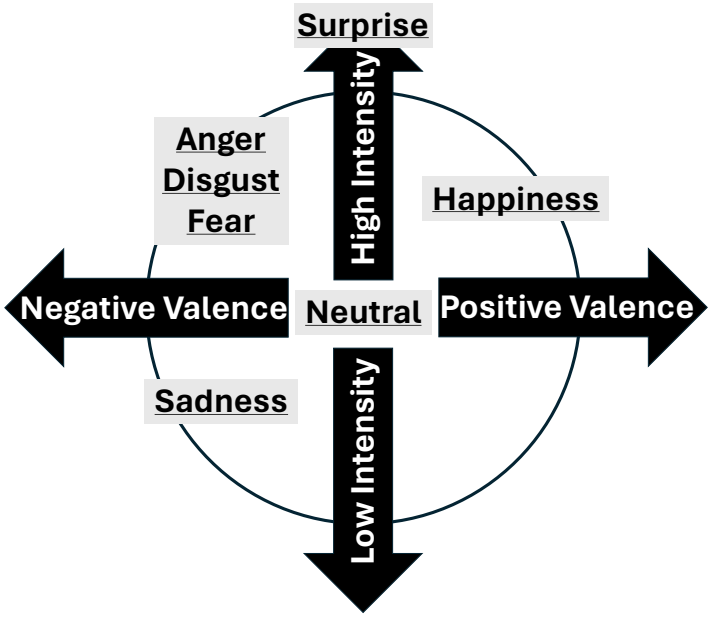


Figure 3.1: Emotion categorical representation adapted in our research, integrating Ekman and Friesen's six basic emotions [24] along with the "Neutral" emotion.

Integrated IMU: Records head movements data from the participants.

3.2.2. Software Components

The emotion collection tool consists of three main software components:

Virtual Environment: Developed with Unity360, and it immerses participants in a controlled setting for presenting video stimuli and performing data annotation after each stimulus.

Researcher Recording System: Comprises the Recording UI and Recording Program, allowing researchers to manage recording sessions, control stimulus playback, initiate recording, and monitor data streams while ensuring synchronized data capture and storage.

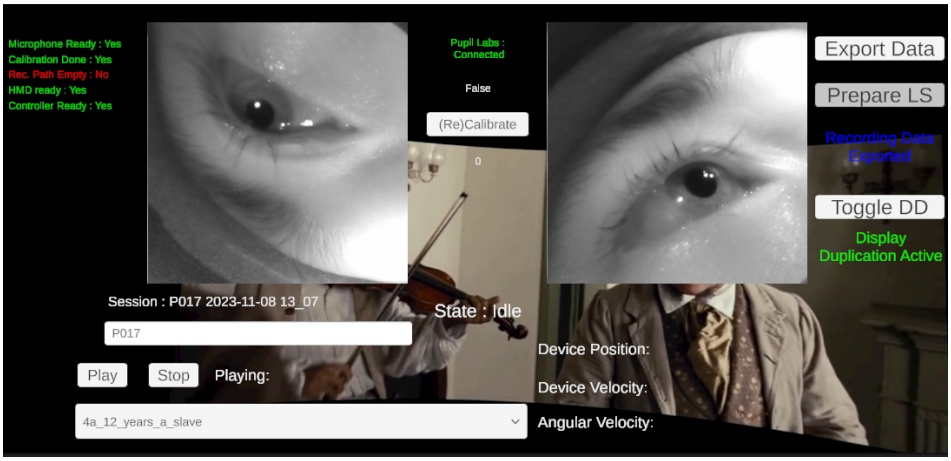
Participant Labeling System: Includes the Labeling UI, enabling participants and researchers to review recorded data, create annotated segments, assign emotion labels and intensity ratings, and compile the final labeled dataset.

Figure 3.2 presents the user interfaces of the Researcher Recording System and the Participant Labeling System. The Researcher Recording System (Figure 3.2(a)) allows the researcher to control and monitor the data collection process, while the Participant Labeling System (Figure 3.2(b)) enables collaborative annotation of the collected data. It is important to note that the Researcher Recording System operates concurrently with the virtual environment, while the participant is immersed and presented with the video stimuli.

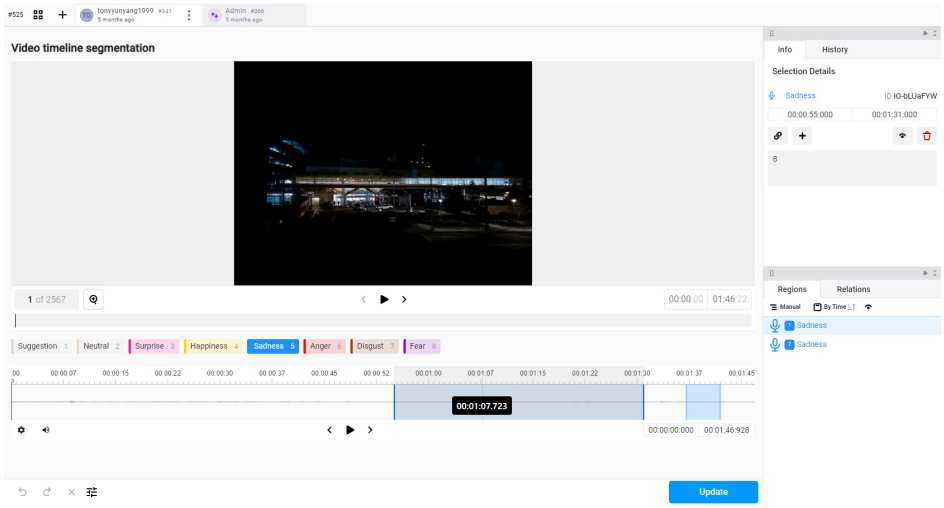
3.3. Dataset Collection Process

The dataset collection process follows a structured approach to ensure consistent and reliable data capture. The data was collected by two researchers: Bishwas, the developer of the tool, gathered data from 18 subjects [74], while 15 more subjects were collected in addition by Tongyun (the author), resulting in a total of 33 subjects. The steps involved in the process are as follows:

1. The participant complete a pre-data questionnaire to gather demographic information and screen for any visual, auditory, or medical conditions that may influence their emotional responses.
2. The participant are instructed to silence their mobile devices to minimize distractions during the data collection process.
3. The researcher sets up the necessary hardware and software components, including Docker, Pupil Capture, SteamVR, and the Unity application. The researcher ensures that all components



(a) Researcher Recording System



(b) Participant Labeling System

Figure 3.2: User interfaces of (a) the Researcher Recording System and (b) the Participant Labeling System. The Researcher Recording System allows the researcher to control and monitor the data collection process, while remaining hidden from the participant who is immersed in the Virtual Environment. The Participant Labeling System enables collaborative annotation of the collected data by both the researcher and the participant.

are functioning correctly and resets the Pupil Capture software to default settings if needed.

4. The researcher measures the participant's interpupillary distance (IPD) and adjusts the head-mounted display (HMD) accordingly to ensure a comfortable and accurate fit.
5. The researcher assists the participant in wearing the HMD and monitors the Researcher Recording System to confirm that all data streams are being captured correctly.
6. The researcher activates the SteamVR Night Mode to minimize visual distractions within the virtual environment.
7. The researcher performs the Unity Gaze Calibration to ensure accurate eye-tracking data collection.
8. The researcher sets the audio volume to a level between 50 and 55 to ensure clear and comfortable audio playback for the participant.
9. The researcher presents the video stimuli to the participant in the sequence specified in Appendix A.1.
10. The researcher follows the process below for each video stimulus, repeating until the last trigger clip:
    (a) The researcher reads the context of the upcoming video to the participant, hinting at the emotion they may potentially experience. This step is omitted for Neutral and Surprise sessions to maintain the integrity of the emotional response.
    (b) The researcher plays the video stimulus for the participant.
    (c) After the video ends, the researcher and participant enter the Participant Labeling System to review the recording together. The researcher asks the participant to label each segment with the emotion they felt and the intensity of that emotion.
11. Once all video stimuli have been presented, the researcher assists the participant in removing the HMD.
12. The participant completes a post-process questionnaire to provide feedback on their experience, and whether felt surprised during the session triggering surprise.
13. The data collection process concludes.

3.4. Data Collection Characteristics

The dataset initially includes a range of modalities, including eye movements, pupil diameter, periocular video recordings, etc. However, though audio and head movement data have been collected, they are excluded from the modalities utilized for the emotion recognition development.

The decision to exclude audio data is because of the nature of the environment where the data collection process was performed. Despite efforts to minimize noise, the presence of other people working in the same space introduce unavoidable background disturbances. While these disturbances do not impact the participants, they introduce irrelevant noise. Furthermore, participants tend to engage in conversations with the researcher throughout the stimuli display. This consistent bias in the audio data limits its usefulness for emotion recognition. Similarly, head movement data are excluded due to the design of the virtual environment. The video clips are not designed for 360° display, as a result, participants remain mostly facing one direction with minimal head movement. This lack of variation in head movement data reduces the informative value for further analysis.

Table 3.1 presents the detailed information of all utilized data. In total, data from 33 subjects have been gathered, out of which 20 are selected for the subsequent stages for the emotion recognition development. The dataset also incorporates self-reported emotion intensity levels from each participant. As mentioned, on the intensity scale, 1 indicates no intensity, 6 represents mild intensity, and 10 signifies very high intensity. An intensity level exceeding 6 is considered indicative of successful emotion elicitation for the corresponding emotion. Figure 3.3 illustrates the distribution of emotion intensity ratings for stimuli across each emotion category. The data reveal a notably high emotion elicitation success rate, averaging above 90% across all emotions. This high success rate underscores the effectiveness of the chosen stimuli in evoking the intended emotions.

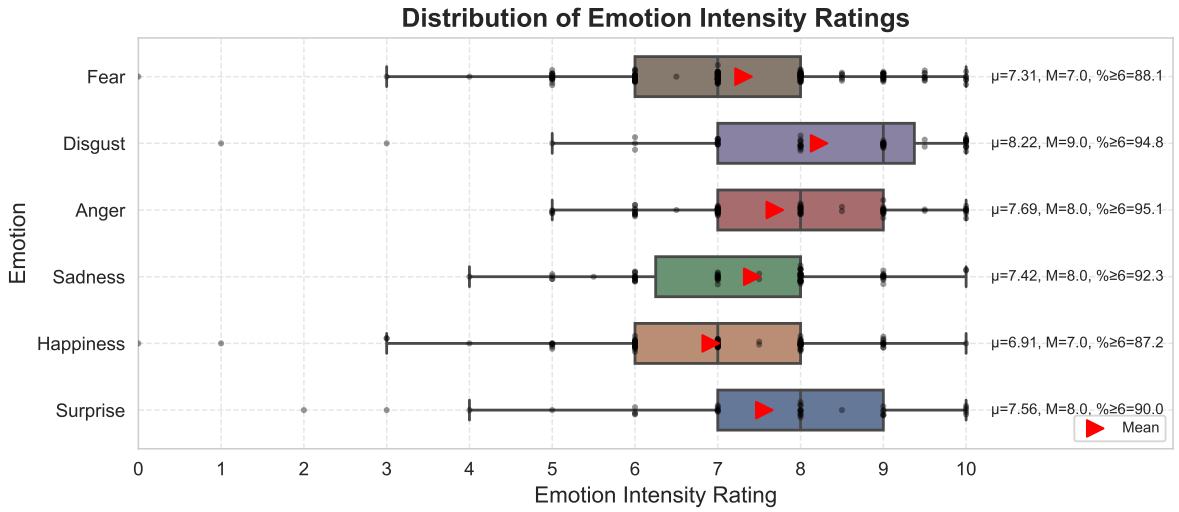


Figure 3.3: Distribution of participants' emotion intensity ratings

Table 3.1: Comprehensive data collection information

Factor	Information
Participant Demographics	
Total Number	20 Participants
Gender Distribution	Female (9), Male (11)
Age Range	Min: 20, Max: 30, Mean: 24.9 ± 2.3
Ethnicity	Caucasian (11), East Asian (4), Middle Eastern (2), South Asian (1), African (1)
Participant Self-Reported States	
Stress Level	Min: 1, Max: 7, Mean: 4.2 ± 1.8 Scale: 1-10 (1: Very Low, 10: Very High)
Fatigue Level	Min: 2, Max: 8, Mean: 4.6 ± 1.7 Scale: 1-10 (1: Not Fatigued, 10: Extremely Fatigued)
Comfort Level	Min: 5, Max: 10, Mean: 7.8 ± 1.3 Scale: 1-10 (1: Very Uncomfortable, 10: Very Comfortable)
Experimental Design	
Elicited Emotions	Neutral, Surprised, Happiness, Sadness, Anger, Disgust, Fear
Stimuli Presentation	14 video stimuli presented to each participant Each emotion elicited by a pair of stimuli ('a' and 'b')
Eye Movement Data	
Gaze Estimation	2-dimensional (x-y coordinates), 240Hz sampling rate
Pupil Diameter	Left and Right eyes, 1-dimensional, 120Hz sampling rate
Video Data	
Periocular Recording	Left and Right eyes, 1 channel (grayscale) each 120fps, 400 x 400 pixels resolution

Note: Sampling rates and recording frame rates are approximates. Pupil diameter sampling rate and periocular recordings frame rates are constrained by the bandwidth between the eye-tracking device and VR headset. Gaze estimation rate is determined by the algorithm used to calculate gaze from pupil data.

4

Data Processing and Preparation

4.1. Raw Data Characteristics

4.1.1. Eye-Tracking System Data Output

The eye-tracking system employed in this study provides a comprehensive set of measurements, including gaze estimation and pupil diameter, as previously outlined in Table 3.1. In addition to these primary measurements, the eye-tracker generates a critical supplementary data, which is the confidence levels associated with each measurement. These confidence levels serve as indicators of the reliability and usability of the corresponding measurements. The confidence levels range from 0 to 1, where 0 denotes complete uncertainty and 1 represents absolute confidence. According to the guidelines provided by the eye-tracker manufacturer, measurements with confidence levels exceeding 0.6 are considered to yield meaningful and reliable information for analysis purposes [19]. Low confidence values are often, but not exclusively, associated with eye blinks or instances where the pupils are not clearly visible to the eye-tracker. Figure 4.1 illustrates this phenomenon, depicting the confidence level plot for both eyes of Subject 2 while viewing clip 0a, designed for a neutral emotional response. Despite the subject performing only 17 blinks throughout the viewing session, the confidence level drops below the 0.6 threshold more than 17 times. Further analysis of the data revealed that 7.39% of left eye data points and 7.70% of right eye data points have confidence levels below 0.6. This discrepancy between blink counts and low confidence occurrences indicates potential reliability issues in the raw eye-tracker output, suggesting that factors beyond eye blinks contribute to data uncertainty.

4.1.2. Challenges Associated with Raw Data

The raw data collected from the eye-tracking system present several challenges that require careful analysis. A primary concern is the variability in confidence levels associated with the measurements. While discarding low-confidence data points might seem intuitive, this approach would disrupt data continuity and create temporal misalignment between different representations, i.e., eye movement data and periocular recordings.

Another significant issue is the inconsistency in sampling rates, as outlined in Table 3.1. The target collection rates for pupil diameter measurements, gaze estimations, and periocular recordings are 120Hz, 240Hz, and 120fps, respectively. However, these rates fluctuate due to bandwidth limitations and the nature of the algorithm to used to derive gaze estimation. This inconsistency manifests in multiple ways: misalignment between left and right eye pupil diameter measurements, discrepancies in the number of pupil measurements relative to periocular recording frames, and an approximate doubling of the gaze estimation sampling rate compared to pupil measurements.

While measurements are sampled at rather high frequencies for better capture of more details, the pupil diameter may not accurately capture human physiological responses. Psychological research indicates that meaningful pupil diameter changes occur at rates just above 9Hz [82], suggesting that higher sampling rates may introduce noise unrelated to actual behavioral reactions.

Furthermore, The high-resolution periocular recordings (400×400 pixels) captured at 120fps impose high computational demands. Processing this data volume on the fly require substantial resources, which can constrain real-time applications.

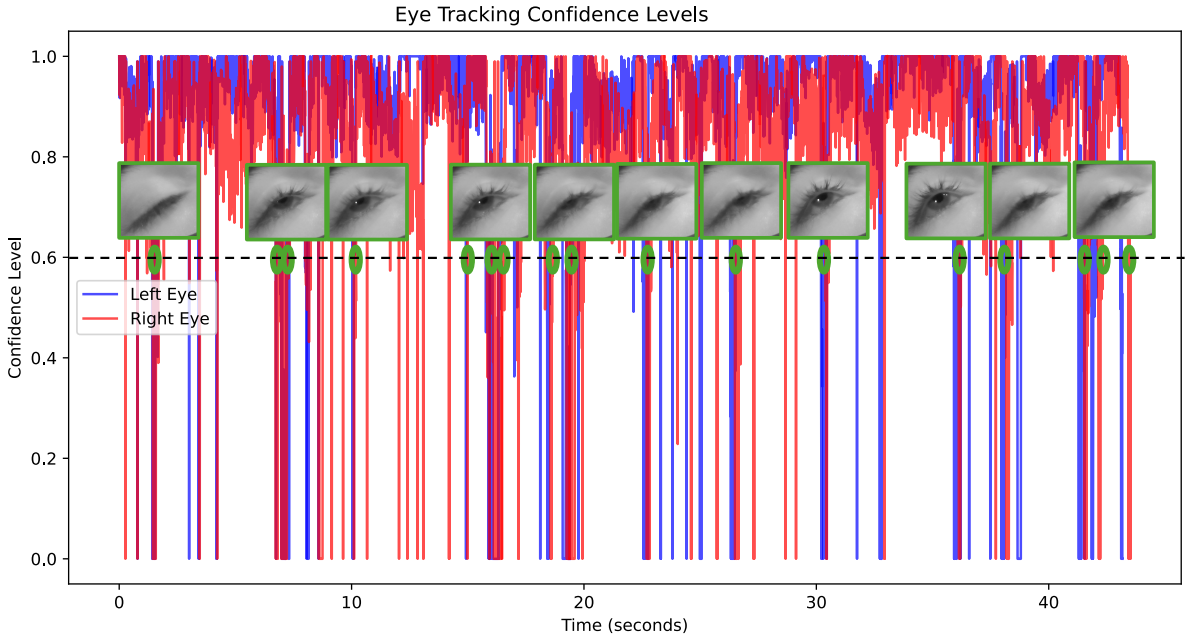


Figure 4.1: Confidence level plot for Subject 2 viewing clip 0a from "Ex Machina" (neutral emotion elicitation). The subject blinked 17 times, yet confidence levels dropped below the 0.6 threshold more frequently. 7.39% and 7.70% of data points had confidence levels below 0.6 for the left and right eyes, respectively.

These challenges underscore the necessity for sophisticated data processing to mitigate noise, synchronize different representations of data, and efficiently handle the computational demands. Addressing these challenges ensures the reliability of the final dataset for emotion recognition.

4.2. Data Processing

This section addresses the challenges of noise mitigation and reduction of micro-computational demands associated with individual samples. The synchronization of different data representations and the reduction of macro-computational demands during the training process are detailed in Section 4.3. Figure 4.2 illustrates the comprehensive data processing workflow. Each filtering step for eye movements is followed by a "Mask and Interpolate" process, which ensures data continuity by interpolating filtered data points rather than directly eliminating them [65]. Data from all representations are scaled to the range of 0 to 1 at last, a standard practice that stabilizes the training process in the further stage of model development by preventing large gradient values from causing significant weight fluctuations [9].

4.2.1. Processing Gaze Estimation

Gaze estimation data are initially filtered based on the confidence level indicated by the eye-tracker (confidence level ≥ 0.6), followed by masking and interpolation of low-confidence data points. Subsequently, Median Absolute Deviation (MAD) with a threshold of 3 is applied to identify extreme outliers. MAD is preferred for outlier detection due to its robustness against extreme values compared to standard deviation-based methods, and its suitability for non-normally distributed data [48]. The mask and interpolate process then follows to address any gaps created by masked outlier data points. Finally, the data are standardized and normalized on a session-specific level, meaning that data from each elicitation session of subjects are standardized and normalized independently [4]. This final step is accounting for slow drifts in sensors, calibration processes, and individual differences in gaze patterns, enabling more accurate comparisons across participants and experimental conditions. Figure 4.3 illustrates the comparison between raw and processed gaze estimation data.

4.2.2. Processing Pupil Diameter Samples

Pupil responses exhibit slower changes compared to eye gaze [82]. This difference requires additional effort to filter pupil measurements to obtain useful information. While the processing pipeline for pupil

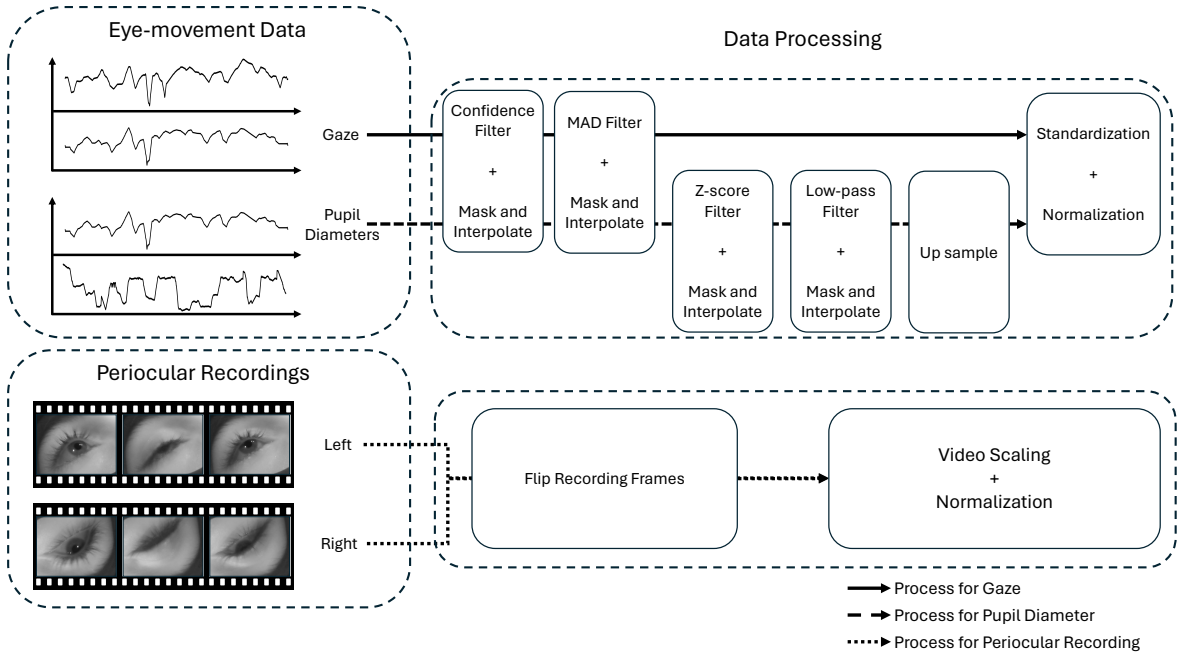


Figure 4.2: Overview of the data processing workflow. 'Mask and Interpolate' refers to the linear interpolation of filtered data.

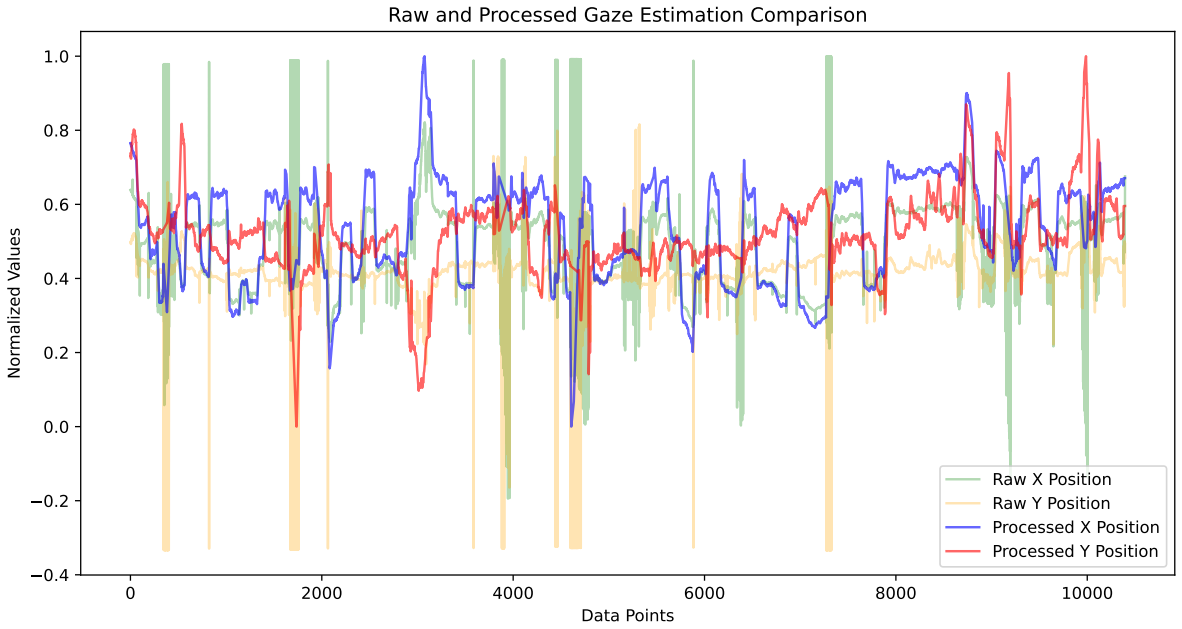


Figure 4.3: Comparison of raw and processed gaze estimation for Subject 2 viewing clip 0a from "Ex Machina" (neutral emotion elicitation).

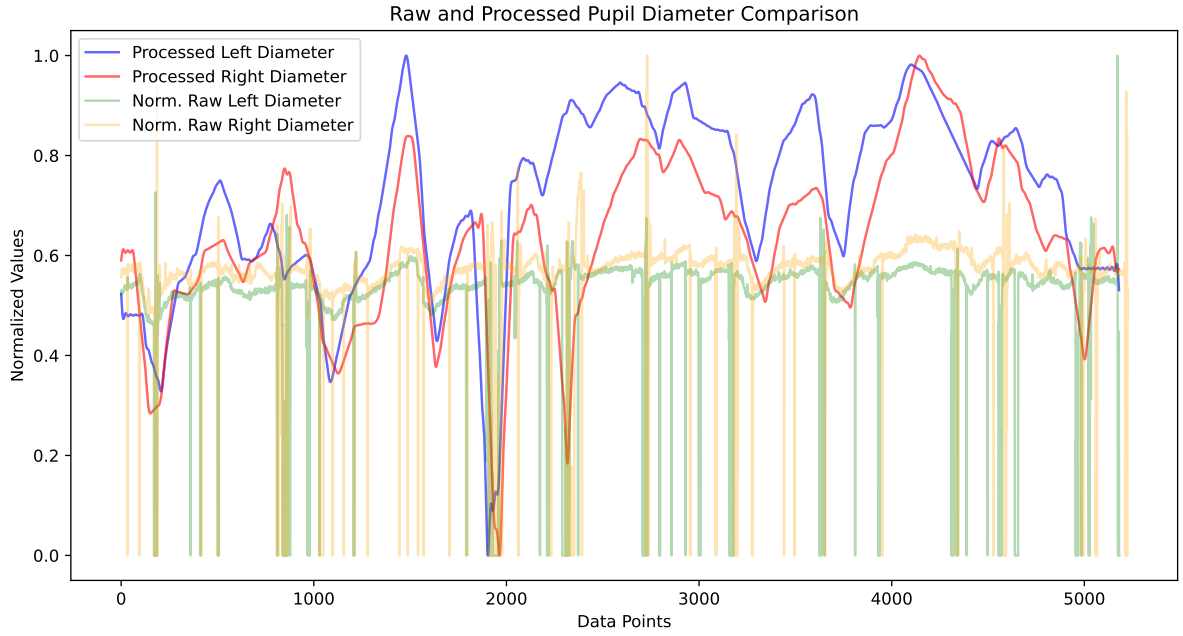


Figure 4.4: Comparison of raw and processed pupil measurements for Subject 2 viewing clip 0a from "Ex Machina" (neutral emotion elicitation).

diameter data shares similarities with that of gaze estimation data, additional steps are included to account for the nature of pupil responses. The initial processing stages mirror those applied to gaze estimation: filtering based on confidence levels, MAD outlier detection (threshold = 3), each followed by masking and interpolation. To address the potential reintroduction of local outliers after interpolation, a multi-pass filtering approach is employed [42]. This approach first applies Z-score filtering (threshold = 0.1) [58], with subsequent masking and interpolation to maintain data continuity. Given the relatively slow nature of pupillary changes, a low-pass filter (cutoff = 5Hz) is then applied to smooth out rapid fluctuations [42, 79]. To ensure the alignment with the gaze data, the pupil measurements are up sampled to 240Hz. Finally, mirroring the last process of gaze estimation data, the pupil diameter data are standardized and normalized. Figure 4.4 shows the comparison of the pupil measurements that are not yet upsampled to 240Hz with normalized raw data, to facilitate comparison with the same amount of data points.

4.2.3. Processing of Periocular Video

The periocular recording video of the right eye is initially processed by flipping both horizontally and vertically. This transformation ensures that the right eye data is in the same orientation as the left eye data, enabling the use of a single eye feature extractor that focuses on extracting general and common changes in the periocular region from both eyes [72]. Following the orientation alignment, the original 400×400 pixel resolution of all video frames are adjusted to 224×224 pixels. This reduction decreases the required computational resources on the fly while maintaining a sufficient level of temporal detail within the videos. Subsequently, all pixel values from both videos are normalized by dividing by 255, a standard practice in image processing [28].

4.3. Dataset Preparation

The effective preparation of the dataset is essential for training the emotion recognition model in the subsequent stage. This process addresses the challenge of macro-computational demands, encompassing two primary aspects: managing the extensive volume of collected data and facilitating efficient data access during the training process.

The collected raw data comprises 276.6GB of information, processing them on the fly would exceed the memory capacity of the training system. Moreover, the raw data consists of continuous data streams, which presents challenges for efficient processing. In order to ensure efficiency and consis-

tency across various training sessions, it is crucial to create a static set of processed samples that can be efficiently accessed during training phase. To address these challenges, a data preparation pipeline is implemented, adhering to the processing steps detailed in Section 4.2. This approach involves processing and systematically organizing the data prior to training, thereby reducing the computational demands. By enabling more efficient access to a consistent set of processed samples, this method not only optimizes training time but also ensures that the same dataset can be utilized reliably across multiple experiments.

Figure 4.5 presents the comprehensive architecture of the dataset preparation pipeline. The pipeline processes raw data from each subject and session independently. Initially, it reads the raw data alongside emotion elicitation times derived from self-reported emotion intensity levels. These data undergo processing, segmentation, alignment, and label assignment before storage in HDF5 (Hierarchical Data Format version 5) format. This process continues sequentially until the final session of the last subject is processed. Subjects are processed in ascending numerical order, with sessions following the sequence 0a, 0b, 1a, 1b, through 6b, comprising 14 sessions per subject. It is noteworthy that sessions where subjects reported no emotion elicitation are excluded from processing. The following subsections provide a details of each component within this pipeline, explaining the methods employed to transform raw data into a format conducive to model training in the further stage of development.

4.3.1. Data Segmentation & Synchronization

The data segmentation and synchronization process involves reading and processing the raw data from each session of each subject, following the procedures outlined in Section 4.2. These procedures result in periocular recordings at 120 fps, gaze estimation at 240 Hz, and pupil diameter measurements at 240 Hz. Concurrently, the emotion elicitation time for each session and subject is retrieved from pre-stored user self-reported intensity levels (detailed in Appendix A.2). The emotion elicitation time is defined as the beginning of the first segment where subjects report an emotional intensity of 6 or higher on the predetermined scale. The emotion elicitation segment extends from this point until the end of the recording, as depicted in Figure 4.6. The elicitation time serves as a reference point for initiating the cropping of different representations.

The cropping process utilizes a specified window size for each experiment. The method begins cropping from the end of each dataset, proceeding backwards until the elicitation time is reached. This approach is based on the intuition that different representations are synchronized at their endpoints, ensuring that cropped samples contain approximately the same information across representations. By initiating the process from the end, the inclusion of relevant data after the user-reported elicitation point is maximized and the risk of losing samples that do not fit within the specified window size is mitigated. While the final sample may extend beyond the emotion elicitation time, it remains valid as long as it contains data claimed to elicit the emotion. This end-aligned cropping strategy ensures consistent temporal alignment across different representations. It is worth noting that the method incorporates the capability for overlapping windows. However, all experiments in this study utilize samples created in a non-overlapping manner. This decision increases the recognition difficulty by reducing the likelihood of recognition based solely on common segments across different samples.

The outcome of this method is arrays of data, each array containing all the samples cropped from representations of the session. These arrays are then prepared for the subsequent process of data storage. A pseudo-code representing this method could be found in Appendix A.3 included with more details.

4.3.2. Efficient Data Storage & Retrieval

Following the segmentation and synchronization of data, the pipeline incorporates an efficient storage and retrieval system. This component addresses the challenges of managing large volumes of raw data and ensuring efficient access during model training.

The system utilizes the HDF5 file format for data storage. HDF5 is chosen for its ability to handle large, complex datasets and its support for high-performance I/O operations. HDF5 allows for the organization of heterogeneous data types within a single file, which is beneficial for storing multi-representation data alongside emotion ground truth labels. The data is structured into separate groups within the HDF5 file for each representation: periocular video frames, gaze coordinates, pupil diameters, and ground truth labels. Each group contains datasets that store the processed samples and their corresponding ground truth, with data and labels corresponding to the same instance stored at

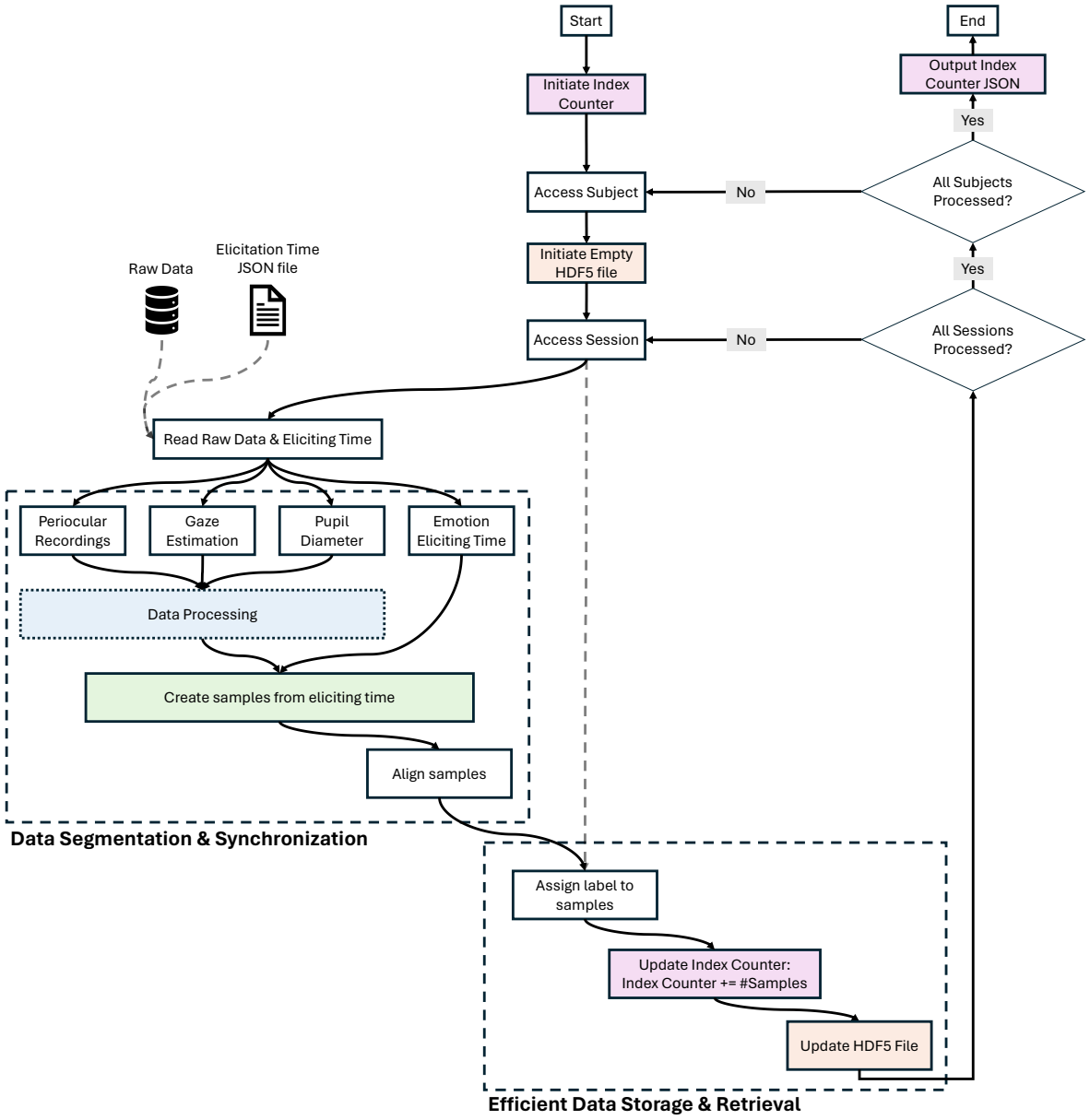


Figure 4.5: Overview of the architecture of the dataset preparation pipeline.

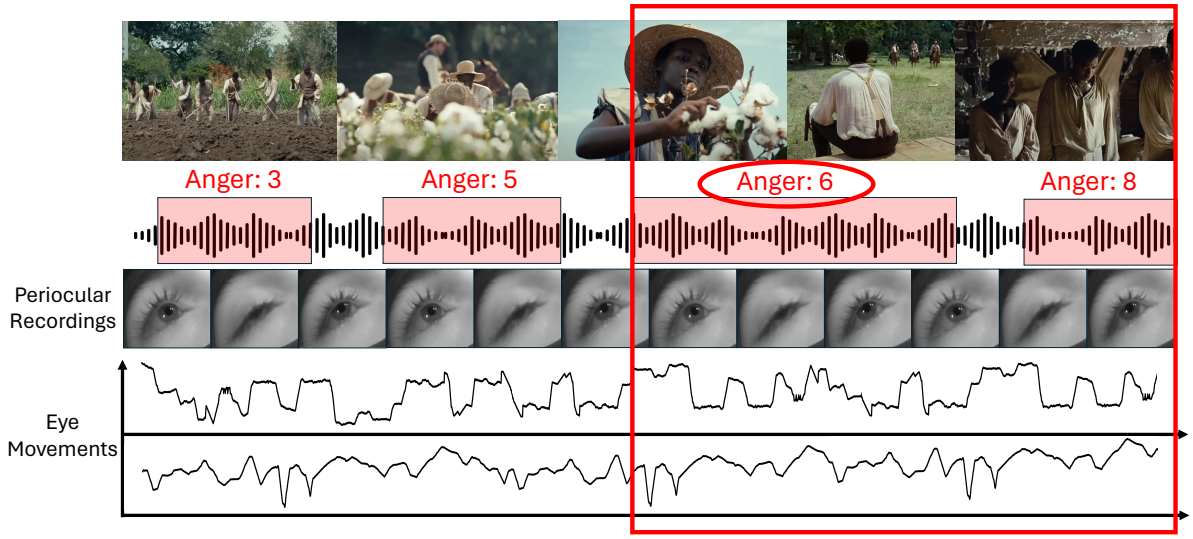


Figure 4.6: Example of the data segmentation, illustrating the identification of emotion elicitation time and subsequent emotion elicitation segment.

the same entry index. The HDF5 format offers lossless compression algorithms, resulting in significant data reduction. The original 276.6GB of raw data is compressed more than 20 times and stored, with periocular recordings down-sampled to 10fps and eye movements sampled at 120Hz.

Another key advantage of using HDF5 for data storage is its ability to access specific portions of the data without loading the entire file, thus enhancing retrieval speed. As illustrated in Figure 4.5, an index counter is initialized at the beginning of the process, and data samples from each subject are stored in separate HDF5 files. The index counter tracks the indices of all samples from each session of each subject. This information is output to a JSON file, which indicates the minimum and maximum indices for each subject, facilitating the identification of the subject associated with a given index. Additionally, the JSON file stores the amount of samples of each session for each subject, making it easy to determine which session a particular sample belongs to. During the training process, entry indices replace the actual data. When specific data is required, it is fetched from the corresponding HDF5 file using the index. This approach addresses the issue latency caused by computing required data on the fly, and minimizing I/O bottlenecks during training.

4.4. Final Dataset Structure and Characteristics

This section describes the characteristic of the dataset created with a non-overlapping window size of one second, periocular recordings at 10fps, and eye movements at 120Hz. It is important to note that different window sizes yield approximately the same number of samples and dataset size, while changes in recording fps result in proportional scaling in size.

4.4.1. Dataset Composition and Distribution

The dataset comprises 15,987 samples from 20 subjects, totaling 13.5GB in size. This represents a significant reduction from the 276.6GB of raw data, with a compression factor of 20.5. Figure 4.7 illustrates the distribution of samples across subjects. The mean sample count per subject is 798.35, with a standard deviation of 65.10. Figure 4.8 depicts the distribution of samples across emotion classes and the two sessions for each emotion. A notable data imbalance exists among different emotions, which can be attributed to the nature of elicitation methods for different emotions. As Paul Ekman argues, most emotions can be repeatedly evoked by stimuli [23], such as maintaining a scary atmosphere or displaying violent scenes to elicit disgust. However, this approach is challenging for surprise, as stimuli that consistently evoke surprise are difficult to design. This imbalance is not unique to the collected dataset in this study, other facial emotion datasets also exhibit over-representation of certain emotions, particularly happiness and other positive emotions, compared to negative emotions [27].

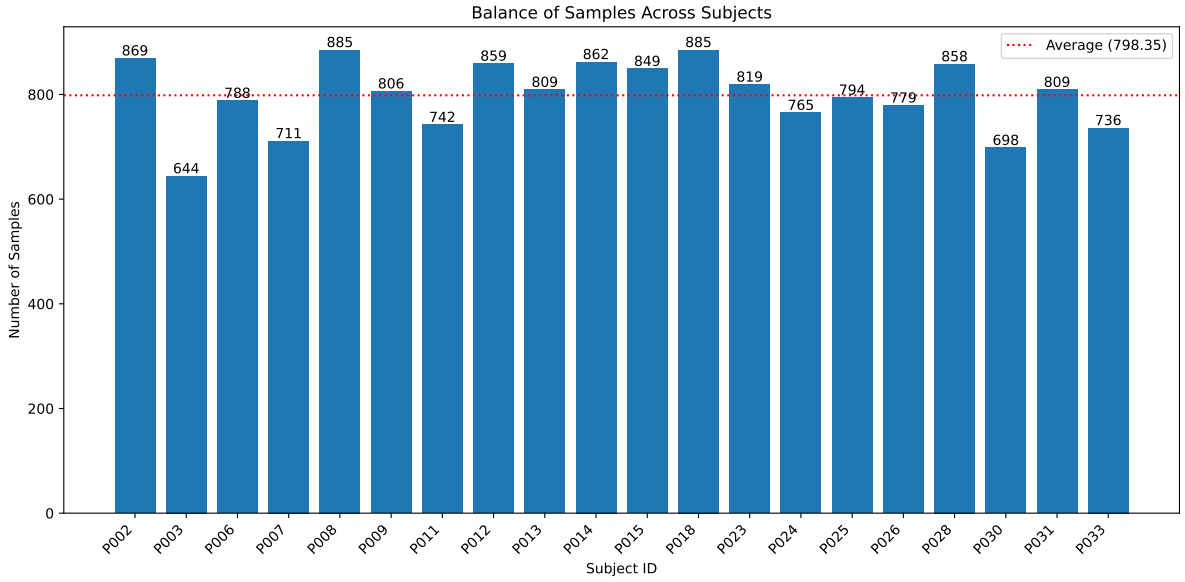


Figure 4.7: Distribution of samples across subjects in the dataset created with a non-overlapping window size of one second, periocular recordings at 10fps, and eye movements at 120Hz

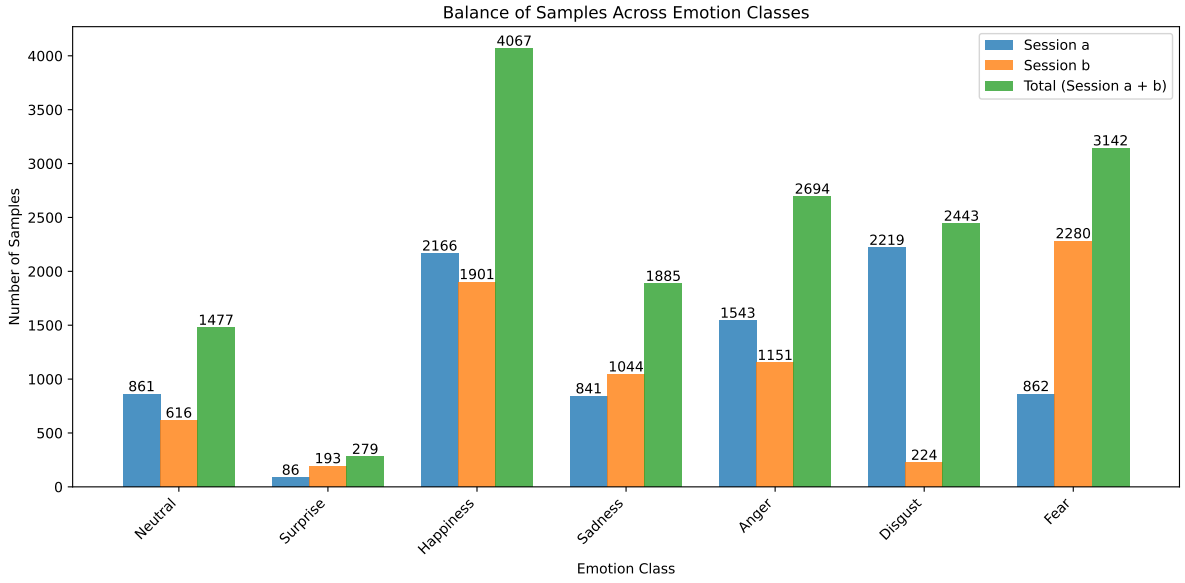


Figure 4.8: Distribution of samples across emotions and sessions in the dataset created with a non-overlapping window size of one second, periocular recordings at 10fps, and eye movements at 120Hz

4.4.2. Data Organization and Characteristics

The dataset is structured with individual subject data stored in separate HDF5 files. Data retrieval is facilitated by indices and an output JSON file, the details of which is presented in Appendix A.4. This JSON file, containing index counts, is generated during the dataset preparation pipeline described in Section 4.3. It includes the minimum and maximum indices for each subject and the sample count for each session. The data retrieval process involves two steps: first, identifying the subject an index belongs to by comparing it to the min and max indices, and second, accessing the data from the corresponding HDF5 file using an adjusted index (calculated as the given index minus the subject's minimum index). A similar method is employed to determine the session for each sample.

The HDF5 files are composed of four primary datasets: Periocular Left, Periocular Right, Eye Movements, and Labels. Their structures are as follows:

Periocular Left and Periocular Right: These datasets contain periocular recordings for the left and right eyes, respectively. Each dataset has a shape of (X, 1, 10, 224, 224) with a float32 data type. X represents the number of samples created with the non-overlapping window, each containing 10 frames of 224×224 pixel gray-scale frames.

Eye Movements: This dataset stores gaze estimation and pupil diameter measurements with a shape of (X, 120, 4) and float32 data type. Each sample is a four-dimensional multivariate time series, where the last dimension contains X-Y coordinates of gaze estimation and pupil diameter measurements for both left and right eyes.

Labels: This dataset encompasses metadata and labels for each sample. It contains X entries, each with a complex data type including subject ID, emotion label, session label, window number, frames per second, frame size, frame channels, and timestamp rate. While the emotion label is primarily used as the ground truth for prediction tasks, the additional metadata is retained to facilitate potential retrieval of specific information of samples when necessary.

5

Emotion Recognition Method

5.1. Emotion Recognition Model Architecture Overview

An emotion recognition method by utilizing only periocular recordings and eye movements is proposed. The model incorporates multiple data representations to enhance the robustness of emotion recognition. It comprises four main components: a periocular feature extractor ($\mathbf{F}_V$), an eye-movement feature extractor ($\mathbf{F}_E$), a feature fusion module, and an emotion classifier. The periocular feature extractor processes periocular recordings from both left and right eyes, while the eye-movement feature extractor analyzes eye movement data. The feature fusion module then integrates the features extracted from each distinct input representation. Finally, the emotion classifier predicts the user's emotional state based on the fused feature. Figure 5.1 presents an overview of the overall architecture.

The periocular recordings are represented as $\tilde{x}_l \in \mathbb{R}^{TWHC}$ and $\tilde{x}_r \in \mathbb{R}^{TWHC}$ for the left and right eyes, respectively. T denotes the temporal dimension (number of frames), while W, H, and C represent the width, height, and channels of each frame. The eye movement data is represented as $\tilde{x}_e \in \mathbb{R}^{wm}$, where w is the sequence length of eye movements, and m is the feature dimension, which in this case is the sum of dimensions of gaze estimation and pupil diameter measurements. The subsequent sections provide detailed descriptions of the processing methodology for each input and their integration to produce the final emotion classification.

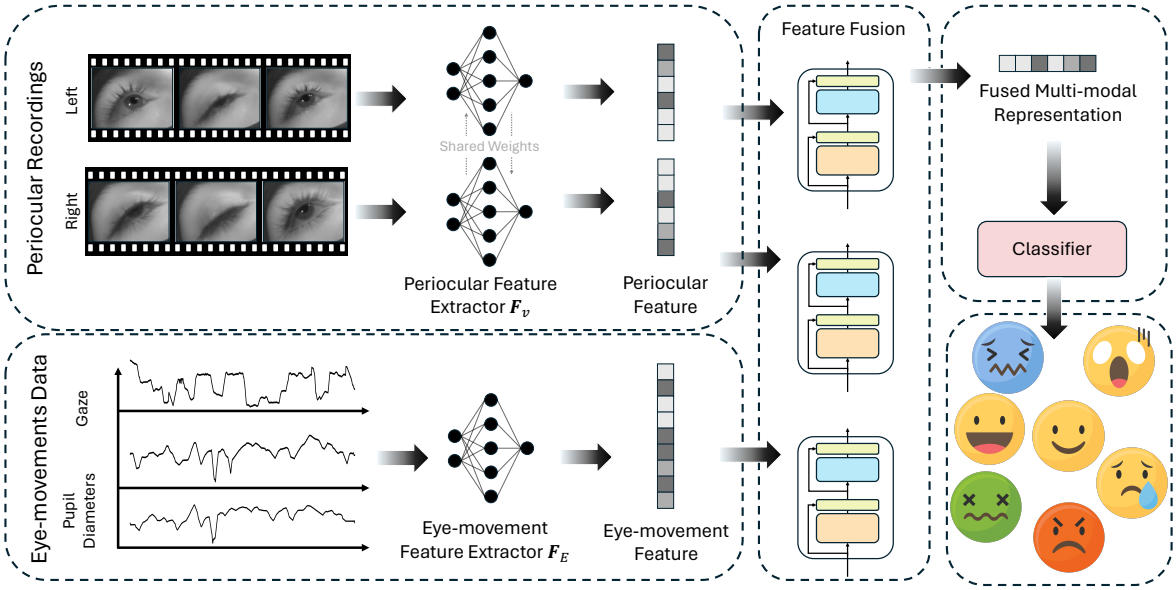


Figure 5.1: Architectural overview of the proposed emotion recognition model.

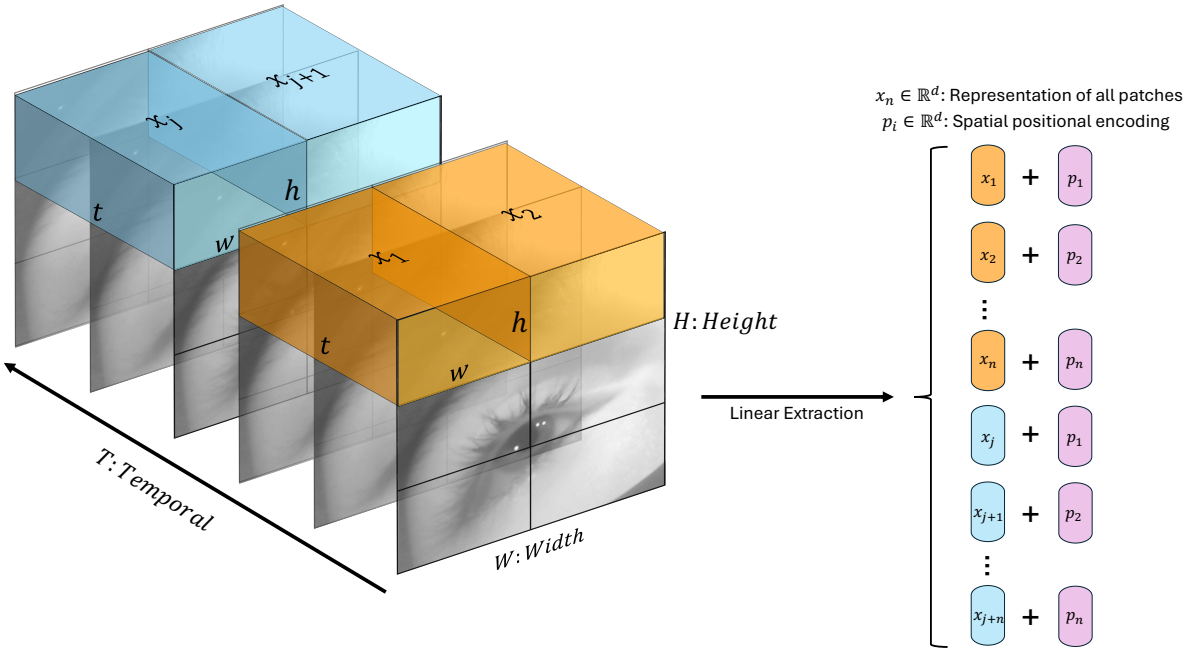


Figure 5.2: Tubelet embedding: Linear extraction of non-overlapping spatial-temporal patches from the input video.

5.2. Periocular Feature Extraction

The periocular feature extractor employs the tubelet embedding method and utilizes the most efficient variant of the Video Vision Transformer (ViViT) model proposed by Arnab et al. [3]. This approach is chosen for its ability to effectively capture spatio-temporal features from video data, demonstrating state-of-the-art performance on video classification tasks.

5.2.1. Tubelet Embedding

Figure 5.2 illustrates the tubelet embedding process, which extracts non-overlapping spatio-temporal patches from the entire video content of a sample. Each patch is then linearly projected into a d-dimensional space to form input tokens for the model. This method allows for efficient processing of video data by preserving both spatial and temporal information in the extracted features.

For each input $\tilde{x}_l \in \mathbb{R}^{TWHC}$ and $\tilde{x}_r \in \mathbb{R}^{TWHC}$, dimensions of each patch are defined as (t, w, h, C). A total of $n_t \times n_w \times n_h$ patches can be extracted from the video, where $n_t = \lfloor \frac{T}{t} \rfloor$, $n_w = \lfloor \frac{W}{w} \rfloor$ and $n_h = \lfloor \frac{H}{h} \rfloor$. To ensure an integer number of patches in each dimension, the input dimensions should be divisible by the corresponding defined patch dimensions. The extracted patches are denoted as $\tilde{x}_{t',w',h'} \in \mathbb{R}^{twhC}$, where $t' = [0, 1, ..., n_t - 1]$, $w' = [0, 1, ..., n_w - 1]$, and $h' = [0, 1, ..., n_h - 1]$. The extracted patches are linearly projected into the d-dimensional space and are augmented with spatial positional embedding to form the sequence of tokens:

$$x'_{t',w',h'} = W_{te}\tilde{x}_{t',w',h'} + b_{te} \tag{5.1}$$

where $x'_{t',w',h'} \in \mathbb{R}^d$, and $W_{te} \in \mathbb{R}^{twhC \times d}$ and $b_{te} \in \mathbb{R}^d$ are learnable parameters.

$$x_{t',w',h'} = x'_{t',w',h'} + p_{w',h'} \tag{5.2}$$

where $x_{t',w',h'} \in \mathbb{R}^d$, and $p_{w',h'} \in \mathbb{R}^d$ represents spatial positional embedding. These embedding are repeated for each frame along the t' dimension n_t times, ensuring that n_t patches of the same temporal index receive the same positional embedding.

5.2.2. Factorized Encoder

The most efficient variant of ViViT, the factorized encoder, is illustrated in Figure 5.3. It consists of spatial and temporal encoders, both sharing the same structure and utilizing pre-norm transformer encoder

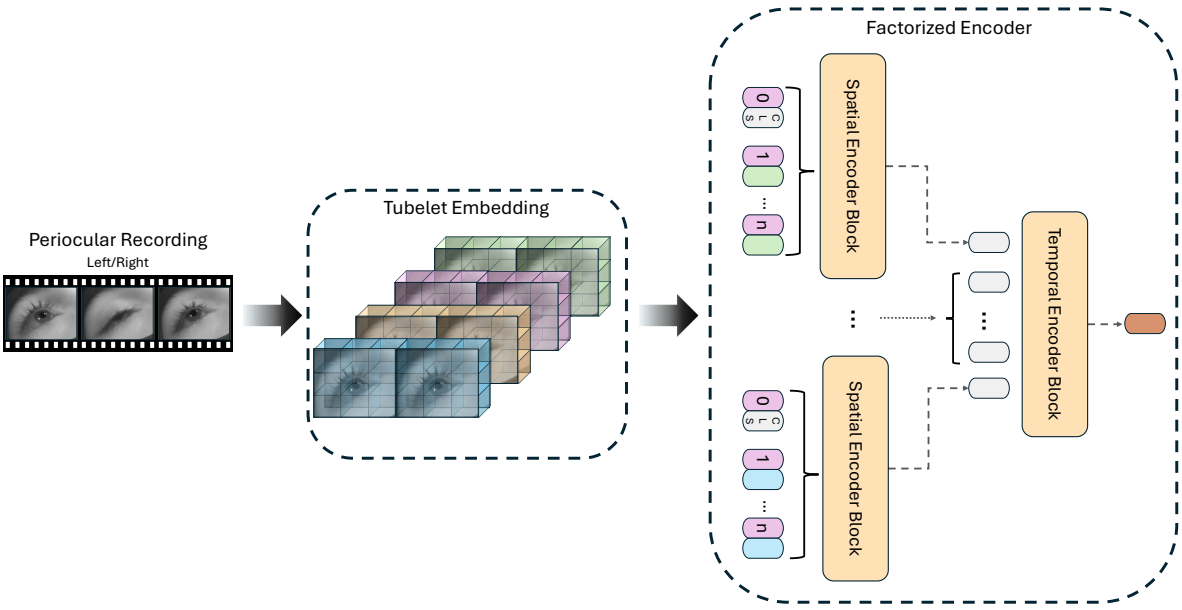


Figure 5.3: Video feature extractor model architecture overview. The model comprises two sequential transformer encoder blocks: the first extracts features from tokens of the same spatial index, while the second facilitates interaction between features from different temporal indices.

blocks (detailed in Appendix A.5).

Spatial Encoder

The tokens formed by the tubelet embedding $x_{t',w',h'} \in \mathbb{R}^d$ are arranged in a multi-dimensional tensor of shape $(n_t, (n_w n_h), d)$, where n_t is the number of temporal indices, $(n_w n_h)$ is the number of spatial tokens per temporal index, and d is the dimension of each token. This arrangement enables the spatial encoder to interact among tokens of the same temporal index. The aim of the spatial encoding process is to extract a single token from spatial tokens of each temporal index for representation. There are two approaches to extract this representation token: one is to attach an extra CLS token to each temporal index, and the other is to perform global pooling of all spatial tokens in each temporal index. Using the CLS approach, a learnable CLS token $x_{cls} \in \mathbb{R}^{1\times1\times d}$ is concatenated to each temporal index, resulting in the multi-dimensional tensor $X_{\texttt{spatial\_input}} \in \mathbb{R}^{n_t(n_w n_h+1)d}$. The spatial encoder then processes this input through multiple transformer encoder blocks:

$$X_{\texttt{spatial\_output}} = \texttt{SpatialEncoder}(\texttt{spatial\_input}) \tag{5.3}$$

The spatial encoder consists of N encoder blocks, where the output of each block serves as the input to the subsequent block:

$$Y_s^n = \texttt{SpatialEncoderBlock}(Y_s^{n-1}), \quad n = 1, \ldots, N \tag{5.4}$$

where Y_s^n is the output of the n-th encoder block, and $Y_s^0 = X_{\texttt{spatial\_input}}$ is the initial input sequence. Each spatial encoder block comprises layer normalization (LN), multi-head self-attention (MSA), and feed-forward network (FF) sublayers:

$$\texttt{SpatialEncoderBlock}(Z) = \texttt{LN}(\texttt{FF}(\texttt{LN}(\texttt{MSA}(\texttt{LN}(Z)))) + Z) + Z \tag{5.5}$$

where Z represents the input to each block. The output of the spatial encoder is $Y_s^N \in \mathbb{R}^{n_t(n_w n_h+1)d}$, and then each of the CLS token is extracted from each temporal index as a representation of the spatial information, resulting in the final output $X_{\texttt{spatial\_output}} \in \mathbb{R}^{n_t d}$.

Temporal Encoder

The final output of the spatial encoder $X_{\texttt{spatial\_output}}$ serves as the input to the temporal encoder, $X_{\texttt{temporal\_input}}$. It undergoes a similar multi-block processing approach as the spatial encoder. Using

the CLS token approach for temporal encoding, a single CLS token is concatenated along the temporal indices, resulting in $X_{\texttt{temporal\_input}} \in \mathbb{R}^{(n_t+1)d}$. The temporal encoder processes this input through multiple transformer encoder blocks:

$$X_{\texttt{temporal\_output}} = \texttt{TemporalEncoder}(X_{\texttt{temporal\_input}}) \tag{5.6}$$

Identical as the spatial encoder, the temporal encoder consists of M encoder blocks:

$$Y_t^m = \texttt{TemporalEncoderBlock}(Y_t^{m-1}), \quad m = 1, \ldots, M \tag{5.7}$$

where Y_t^m is the output of the m-th encoder block, and $Y_t^0 = X_{\texttt{temporal\_input}}$ is the initial input sequence. Each temporal encoder block has the same structure as the spatial encoder block:

$$\texttt{TemporalEncoderBlock}(Z) = \texttt{LN}(\texttt{FF}(\texttt{LN}(\texttt{MSA}(\texttt{LN}(Z)))) + Z) + Z \tag{5.8}$$

where Z represents the input to each block. The output of the temporal encoder is $Y_t^M \in \mathbb{R}^{(n_t+1)d}$, and then the CLS token is extractedas a representation of the whole video sample, resulting in the final output $X_{\texttt{temporal\_output}} \in \mathbb{R}^d$. $X_{\texttt{temporal\_output}}$ is also the final output of the periocular feature extractor.

Multi-Head Self-Attention

Both spatial and temporal encoders use multi-head self-attention mechanism (MSA). For an input $Z \in \mathbb{R}^{l \times d}$, where l is the amount of tokens, the multi-head self-attention is computed as:

$$\texttt{MSA}(Z) = \texttt{Concat}(\texttt{head}_1, ..., \texttt{head}_c)W^O \tag{5.9}$$

where c is the number of attention heads, and $W^O \in \mathbb{R}^{cd_k \times d}$ is the output projection matrix. Each attention head is computed as:

$$\texttt{head}_i = \texttt{Attention}(ZW_i^Q, ZW_i^K, ZW_i^V) \tag{5.10}$$

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.11}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$ are learned parameter matrices, and $d_k = d/c$ is the dimension of each head.

Feed-Forward Network

The feed-forward network in each encoder block is defined as:

$$\texttt{FF}(Z) = W_2(\texttt{GELU}(W_1 Z + b_1)) + b_2 \tag{5.12}$$

where $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$, $b_1 \in \mathbb{R}^{d_{ff}}$, and $b_2 \in \mathbb{R}^d$ are learnable parameters, and d_{ff} is the hidden layer dimension. The network projects the input to a higher dimension and back, and introducing non-linearity through the GELU activation function. it enhances the model's ability to capture complex patterns in the data.

5.3. Eye-Movement Feature Extraction

The eye-movement feature extractor is based on a multivariate time series transformer framework [98]. This approach is chosen for its demonstrated efficacy in processing multivariate time series data, which is crucial for capturing the temporal dynamics in eye movements. The core structure of the framework is adopted, specifically a transformer encoder block in the post-norm fashion (detailed in Appendix A.5). Figure 5.4 illustrates an overview of the adopted architecture.

As mentioned earlier, the eye movements are represented as a multivariate time series $\tilde{x}_e \in \mathbb{R}^{wm}$, where w is the sequence length and m is the feature dimension. Each time step in the sequence is represented by a feature vector $x_{st} \in \mathbb{R}^m$.

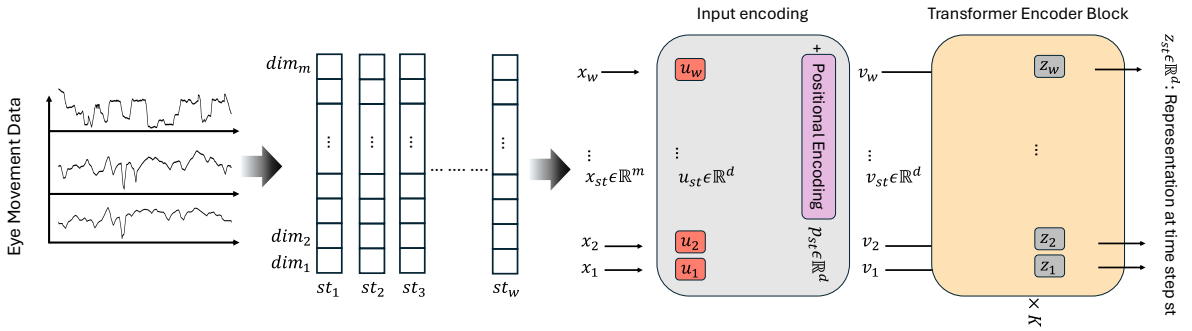


Figure 5.4: Architecture overview of the time series feature extractor for eye movements.

5.3.1. Input Embedding

The embedding process includes projection to a d'-dimensional space and the addition of positional encoding. These steps can be formulated as:

$$u_{st} = W_p x_{st} + b_p \tag{5.13}$$

$$v_{st} = u_{st} + p_{st} \tag{5.14}$$

where $W_p \in \mathbb{R}^{d' \times m}$ and $b_p \in \mathbb{R}^{d'}$ are learnable parameters, and $p_{st} \in \mathbb{R}^{d'}, st \in \{0, ..., w-1\}$ are the positional encoding for each time step st. The resulting sequence, $V \in \mathbb{R}^{wd'} = [v_0, v_1, ..., v_{w-1}]$, serves as input tokens to the transformer encoder.

5.3.2. Transformer Encoder

The transformer encoder for eye movements shares the fundamental structure with the periocular feature extractor (described in Section 5.2), but incorporates adaptations for time series data. Unlike the pre-norm setup used previously, this encoder applies normalization after each component (self-attention and feed-forward network). The post-norm configuration that has shown empirical benefits in time series processing [98]. Additionally, batch normalization, applied across both batch samples and time steps, replaces layer normalization as previously used.

The transformer encoder for eye movements can be represented as a series of encoder blocks:

$$Y_e^k = \texttt{TransformerEncoderBlock}(Y_e^{k-1}), \quad k = 1, \ldots, K \tag{5.15}$$

where Y_e^k is the output of the k-th encoder block, and $Y_e^0 = V$ is the initial input sequence. Each encoder block is defined as:

$$\texttt{TransformerEncoderBlock}(Z) = \texttt{BN}(\texttt{FF}(\texttt{BN}(\texttt{MSA}(Z) + Z)) + Z) \tag{5.16}$$

where Z represents the input to each block, and BN is batch normalization, which replaced layer normalization. The multi-head attention mechanism and the feed-forward network remain identical to those described previously in Section 5.2. The output of the transformer encoder is $Y_e^K \in \mathbb{R}^{w \times d'}$. The final feature representation is then obtained by applying a GELU activation function to the output, followed by mean pooling across the temporal dimension. This process can be summarized as:

$$f_e = \texttt{MeanPool}(\texttt{GELU}(Y_e^K)) \tag{5.17}$$

where $f_e \in \mathbb{R}^{d'}$ is the final output feature vector. The GELU activation introduces non-linearity, while the mean pooling operation reduces the sequence dimension, resulting in a compact representation of the entire eye movement sequence that captures global temporal dependencies.

5.4. Multi-Representation Feature Fusion

The proposed emotion recognition model leverages features extracted from both periocular recordings and eye movements. This section details the fusion approach used to combine these multi-representation features effectively. Figure 5.5 illustrates an overview of the fusion process.

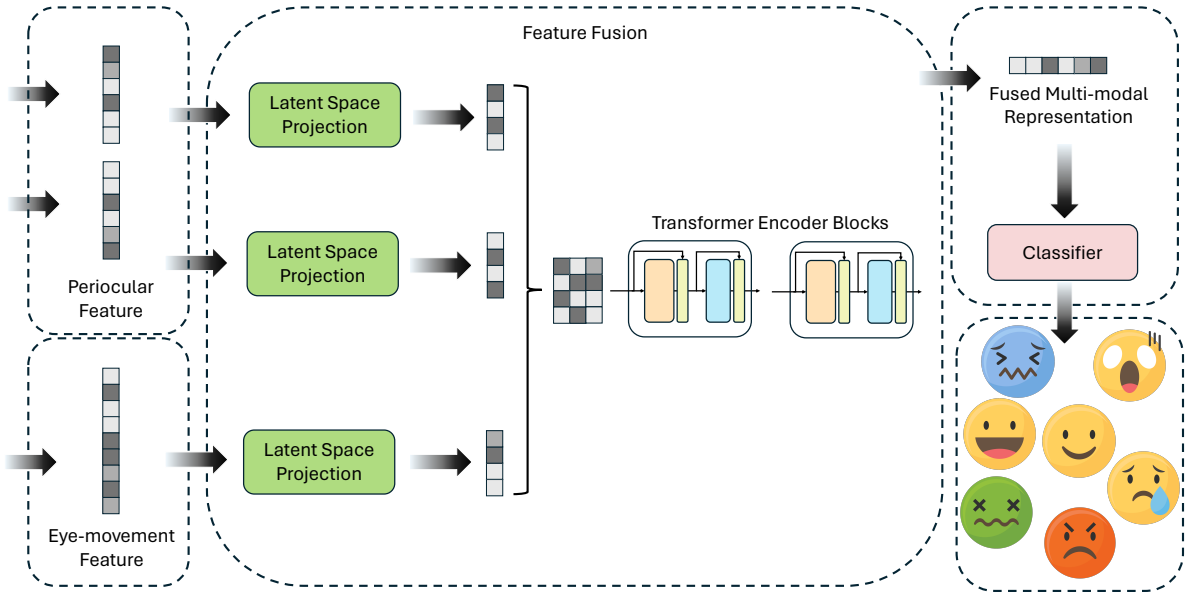


Figure 5.5: Overview of multi-representation cross-attention feature fusion process.

To effectively combine the extracted features from multiple representations, a multi-representation cross-attention fusion approach is proposed. This approach allows the model to learn the internal relations among different features and the relative importance of each. The fusion method consists of three main steps: feature projection, cross-attention processing, and weighted pooling.

5.4.1. Extracted Features

As a result of the feature extraction processes described in the previous sections, the following feature representations are obtained:

$$f_{v,l} = \mathbf{F}_V(\tilde{x}_l) \tag{5.18}$$

$$f_{v,r} = \mathbf{F}_V(\tilde{x}_r) \tag{5.19}$$

$$f_e = \mathbf{F}_E(\tilde{x}_e) \tag{5.20}$$

where $f_{v,l}$, $f_{v,r}$, and f_e correspond to features obtained from the left and right periocular recordings, and the eye movements, respectively.

5.4.2. Cross-Attention Based Fusion

The cross-attention based fusion process involves projecting obtained features from different representations into a same dimension, processing representation features with cross-attention and forming the final fused representation with weighted pooling at last.

Feature Projection

The fusion method begins by projecting each representation's features into a common h'-dimensional space:

$$f_{ri} = W_{ri}x_{ri} + b_{ri}, \quad ri \in \{f_{v,l}, f_{v,r}, f_e\} \tag{5.21}$$

where $W_{ri} \in \mathbb{R}^{h' \times d_{ri}}$ and $b_{ri} \in \mathbb{R}^{h'}$ are learnable parameters, and d_{ri} is the original dimension of representation ri. This step results in dimension-aligned features $f'_{v,l}$, $f'_{v,r}$, and f'_e from all three representations.

Cross-Attention Processing

The aligned features in dimension are then stacked and processed through a transformer encoder in post-norm fashion, which facilitates effective interaction between different representations:

$$F = \texttt{FF}(\texttt{MSA}([f'_{v,l}, f'_{v,r}, f'_e])) \tag{5.22}$$

The output $F \in \mathbb{R}^{n_f \times h'}$ contains refined features that have incorporated information from all representations. In the case where all three representations are involved in the feature fusion process, $n_f = 3$.

Weighted Pooling

Finally, a learnable weighted sum is employed to pool the refined features into a single representation:

$$f_{\texttt{fused}} = \sum_{a=1}^{n_f} \beta_a F_a \tag{5.23}$$

where β_a are learnable pooling weights normalized through a soft-max function, and the final fused representation $f_{\texttt{fused}} \in \mathbb{R}^{h'}$ encapsulates information from all input modalities.

5.5. Emotion Classification

The final stage of the emotion recognition model involves classifying the fused features into discrete emotion categories. The feature vector resulting from the fusion approach serves as input for emotion classification. A simple linear layer is implemented for this task, which can be expressed as:

$$y = W f_{\texttt{fused}} + b \tag{5.24}$$

where $W \in \mathbb{R}^{h' \times cl}$ and $b \in \mathbb{R}^{cl}$ are learnable parameters, cl represents the number of emotion classes, and h' denotes the dimension of the fused representation. This linear transformation maps the fused features to a cl-dimensional space, corresponding to the emotion classes. To transform the linear layer output into class probabilities, the soft-max function is applied:

$$\hat{y}_u = \frac{\texttt{exp}(y_u)}{\sum_{j=1}^{cl} \texttt{exp}(y_j)} \tag{5.25}$$

where $\hat{y}_u$ represents the predicted probability of the u-th class, and y_u is the u-th element of the linear layer output for a given sample. The emotion class with the highest probability is taken as the recognized emotion. For model training, the cross-entropy loss function is employed. Given a batch of B samples, the calculated class probabilities $\hat{y}$, and the true class labels $\tilde{y}$, the loss is computed as:

$$\mathcal{L} = -\frac{1}{B} \sum_{n=1}^{B} \sum_{i=1}^{cl} \tilde{y}_{n,i} \texttt{log}(\hat{y}_{n,i}) \tag{5.26}$$

where it aims to minimize the discrepancy between the predicted probabilities and the true class labels during the training process.

5.6. Model Configuration

This section elaborates on the specific configurations of all building modules of the model described previously. These configurations remain consistent across different experimental setups in subsequent stages of experimentation. The primary variations in the experiments involve the inclusion or exclusion of specific modules, depending on the method being evaluated. This approach facilitates a systematic comparison of different representations and their combinations in emotion recognition tasks while maintaining a consistent architectural design.

The periocular feature extraction module, based on the ViViT architecture detailed in Section 5.2, is designed to extract spatio-temporal features from periocular recordings efficiently. The tubelet embedding first processes video frames in patches of 16×16 pixels spatially and five frames temporally, then projecting the patches to a dimension of 256. This dimension serves as the size for the input tokens and internal representations throughout the module. The architecture comprises three spatial transformer blocks followed by one temporal transformer block, each utilizing eight attention heads. The feed-forward network within each transformer block is designed with a dimension of 1024, four times the dimension of the internal representation. To enhance the model's generalization capabilities and mitigate over-fitting, dropout rates of 0.1 are incorporated in the tubelet embedding, spatial encoder, and temporal encoder.

The eye movement feature extraction module, implemented as a time series transformer, as described in Section 5.3, is designed to process four-dimensional feature vectors representing gaze estimation and pupil diameter measures in eye movement data. This module operates with input tokens of dimension 64 and employs eight attention heads within a single transformer block. The feed-forward network in this module has a dimension of 256, maintaining the four-to-one ratio as in the ViViT module. It also incorporates a dropout rate of 0.1 for regularization in both embedding and encoder components.

In methods incorporating feature fusion, as described in Section 5.4, the extracted feature representations from modules undergo further processing. These representations are projected into a common 256-dimensional space before combination, ensuring that features from different representations are aligned in dimension. The subsequent fusion process utilizes a single transformer block with eight attention heads. The feed-forward network in this module has a dimension of 1024, maintaining the four-to-one ratio. A learnable weighted sum is employed which enables the model to leverage complementary information from different representation features.

6

Experiment Methods and Results

6.1. Experiment Design

6.1.1. Dataset Preparation

Posterior to data processing, emotion-elicited segments are extracted based on user self-reported emotion intensity levels and prepared into datasets, as detailed in Section 4.3. However, the method of extracting segments provided by each user has limitations, as new incoming users' labels are not available in real-world applications. To address this, an alternative method is proposed.

The alternative method provides an estimation of each emotion elicitation time for each user using a 95% confidence interval, derived from the data of all other users. The lower bound calculated for each session serves as the estimated elicitation time. Subsequently, the segment from the determined starting point to the end is extracted and prepared into a dataset using the pipeline detailed in Section 4.3. To assess the accuracy of this estimation, the mean absolute error (MAE) between the estimated and user-provided emotion elicitation times across all sessions is calculated. The average MAE is 6.60 seconds with a standard deviation of 15.47 seconds, indicating that most users experience emotion elicitation at approximately similar points in time.

Datasets using both emotion labels are prepared with recording segments sampled at a maximum of 10fps, and eye movements at 120Hz, utilizing non-overlapping sliding windows with a maximum size of two seconds.

6.1.2. Evaluation Protocol

The experiment employs a 5-fold cross-validation approach across 20 subjects, implementing two primary testing schemes:

Pre-train Testing. This scheme evaluates the model on samples from four unseen subjects in each fold without any fine-tuning. It assesses the model's performance on entirely new subjects after solely pre-training with data from all other sixteen subjects. This approach tests the ability of the model to generalize across different individuals.

Fine-tune Testing. This scheme involves fine-tuning the model with a small proportion of samples from an unseen subject before testing on the remaining samples. This mimics real-world application scenarios where a limited amount of data from a new user is available for the adaptation purpose. This approach evaluates the model's ability to quickly adapt to new users with minimal additional data.

6.1.3. Implementation Details

In the 5-fold cross-validation pre-training process, all samples from the four unseen subjects in each fold are used as testing samples, while all samples from the remaining sixteen subjects are used for training. In the personal adaptation fine-tuning process, the model is fine-tuned on each independent subject. The initial samples of each session from a subject are selected, aligning with the nature of watching videos. This approach simulates a realistic scenario where early data from a new user becomes available for model adaptation. To comprehensively verify the model's performance, its ability to recognize emotions triggered by watching unseen data of the same content (same-session) and its

ability to recognize emotions triggered by watching unseen different content (cross-session) are both evaluated. The implementations of these two distinct criteria are as below:

Same-session. A certain amount of initial samples are taken from both sessions of each emotion and are used as training data, while the remaining samples serve as testing data. This approach assesses the model's ability to recognize emotions triggered by watching unseen data of the same content.

Cross-session. A certain amount of initial samples are taken from one of the sessions of each emotion and are used as training data, and the data of the other session serve as testing data. This process is repeated twice for each subject, as each emotion has two sessions per subject, and the performance is averaged. This approach evaluates the model's ability to recognize emotions triggered by watching unseen different content, testing its ability to generalize across varying stimuli.

6.1.4. Evaluation Metrics

For all experiments, the model's performance is reported as emotion recognition accuracy using the weighted F1-score, which accounts for class imbalance. The weighted F1-score is calculated as follows:

$$\texttt{Weighted F1-score} = \sum_{i=1}^{n} w_i \cdot \texttt{F1}_i \tag{6.1}$$

where n is the number of classes, w_i is the proportion of true instances for class i, and $F1_i$ is the F1-score for class i. The F1-score for each class is the harmonic mean of precision and recall:

$$\texttt{F1}_i = 2 \cdot \frac{\texttt{precision}_i \cdot \texttt{recall}_i}{\texttt{precision}_i + \texttt{recall}_i} \tag{6.2}$$

where precision$_i$ and recall$_i$ are defined as:

$$\texttt{precision}_i = \frac{\texttt{true positives}_i}{\texttt{true positives}_i + \texttt{false positives}_i} \tag{6.3}$$

$$\texttt{recall}_i = \frac{\texttt{true positives}_i}{\texttt{true positives}_i + \texttt{false negatives}_i} \tag{6.4}$$

6.1.5. Baseline Methods

To demonstrate the effectiveness of the proposed approach, it is compared with three other baselines:

Periocular Only. This approach utilizes only the periocular recordings from both eyes as input. It employs the upper half of the architecture shown in Figure 5.1, including the periocular feature extractor to extract features from both eyes and the feature fusion module to combine features extracted from them. This baseline assesses the contribution of visual information from the periocular region to emotion recognition.

Eye Movements Only. This approach uses only the eye movement data as input. It utilizes the lower half of the architecture shown in Figure 5.1, solely including the eye movement feature extractor, and excluding the feature fusion module. This baseline evaluates the performance contribution of eye movement patterns to emotion recognition.

Content Attention. This approach incorporates the recording of the user's view (i.e., the content being watched) along with an attention map. For each frame of the view, an attention heat map is calculated based on the average gaze point of the user on the displayed content. This heat map is then appended to each frame as a fourth channel, as illustrated in Figure 6.1. The dataset is created with the pipeline detailed in Section 4.3. The method employs a single periocular feature extractor of the proposed architecture, which in this case acts as a feature extractor of the content attention, excluding the feature fusion module and eye movement processing components. This baseline presents the performance of combining visual content with gaze information for emotion recognition.

6.2. Performance Evaluation

6.2.1. Model Training

The training process for each evaluated model is designed to optimize performance and ensure convergence. The process encompasses 25 epochs, employing the Adam optimizer with a learning rate

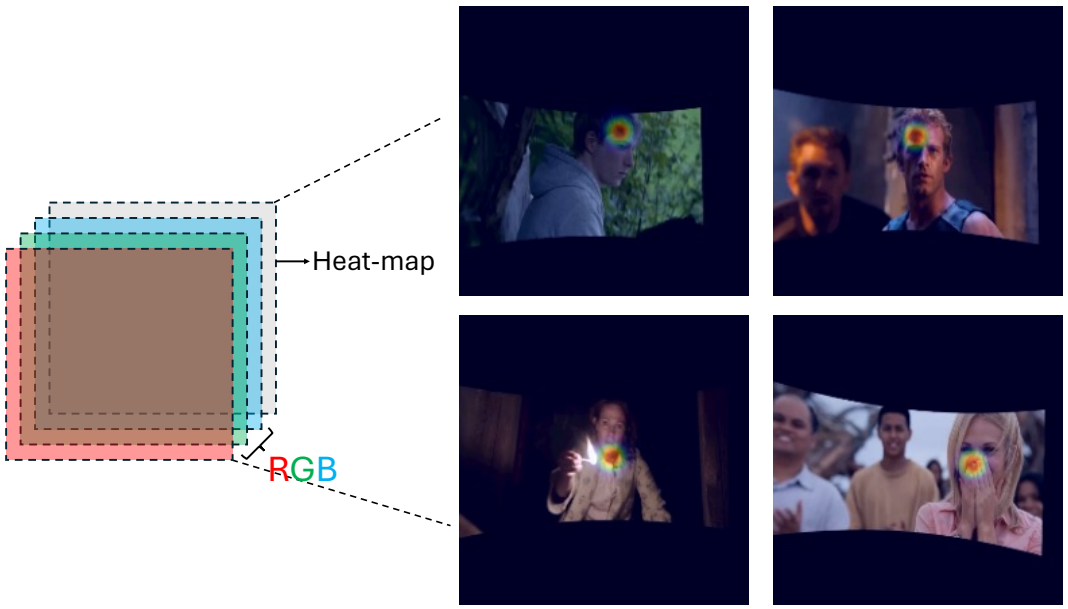


Figure 6.1: Illustration of the content attention heat map overlay on user view frames.

scheduler. The scheduler incorporats a five-epoch warm-up phase, peaking at a learning rate of 8e-5, followed by 20 epochs of cosine annealing, decreasing to 8e-6. The final five epochs maintain this lower learning rate to ensure convergence. It has been found that applying a label smoothing factor of 0.4 improves performance. The intuition to this cause is that label smoothing helps reduce the model's confidence in its predictions, potentially leading to better generalization during pre-training. Consistently, 30% of the training data is allocated for validation, and the entire dataset of unseen subjects is reserved for testing.

For the fine-tuning phase, the same optimizer and learning rate scheduler are utilized, but with modifications to suit the task of adaptation to new subjects. The duration of each phase is extended to five times that of the pre-training, and the peak learning rate is reduced to 1e-5 and the final learning rate to 1e-6, allowing for more gradual adaptation. Label smoothing is removed during fine-tuning to allow the model to capture subject-specific nuances more precisely.

The training process above applies to the proposed method and the periocular-only baseline. However, for the content attention baseline, a different approach is necessary. This baseline employs the same fine-tuning setting but skips the pre-train phase. The decision is made because pre-training is not a reasonable task for the content attention baseline, as it inherently relates content to certain emotions. Moreover, given that the stimuli presented to each participant are the same, including a pre-training phase would not provide a fair comparison with the other methods. The eye movements-only baseline requires a modified training approach. It is trained for a greater number of epochs with a larger learning rate to ensure convergence.

6.2.2. Pre-training Performance Analysis

The pre-training performance of the proposed model is evaluated on samples from unseen subjects without personalized fine-tuning adaptation. This evaluation aims to assess the model's generalizability across different individuals. Table 6.1 presents the performance results across various window sizes for different representation combinations. The data consistently demonstrates that the combination of eye movements and periocular data outperforms single-representation approaches, with the best performance at the two-second window size (F1-score of 0.52). This multi-representation approach shows a 15.6% improvement over the periocular-only method (F1-score of 0.45) and a significant 73.3% improvement over the eye movements-only method (F1-score of 0.30) at the same two-second window size. The superior performance of the multi-representation approach can be attributed to its ability to capture complementary information from both eye movements and periocular recordings. Eye movements provide dynamic temporal information about gaze patterns, while periocular recordings offers

Table 6.1: Pre-training performance across various window sizes

Modalities	Window Size (s)			
	0.5	1.0	1.5	2.0
Eye Movements	0.26	0.27	0.29	0.30
Periocular	0.45	0.44	0.46	0.45
Eye Movements + Periocular	0.43	0.44	0.46	**0.52**

Table 6.2: Fine-tuning performance with one-second window: Comparison across modalities

Modalities	10% Same-session	10% Cross-session
Eye Movements	0.46	0.23
Content Attention	0.64	0.17
Periocular	0.82	0.54
Eye Movement + Periocular	**0.84**	**0.70**

visual cues from the eye and muscle movements.

However, an exception to this trend occurs at the 0.5-second window size, where the proposed approach slightly under-performs (F1-score of 0.43) compared to the periocular-only method (F1-score of 0.45). This may be attributed to the limited information providing by eye movements in such a short time frame. In this case, the eye movement data might introduce noise rather than valuable features to the model, potentially obscuring the more reliable periocular information. This observation suggests the existence of a minimum temporal threshold for effectively leveraging eye movement data in emotion recognition tasks. At shorter time scales, the patterns of eye movements may be less informative for emotion recognition. As the window size increases, the eye movement data becomes more meaningful, contributing positively to the overall performance of the multi-representation model.

6.2.3. Fine-tuning Performance Analysis

The fine-tuning performance analysis evaluates the effectiveness of personal adaptation based on the pre-trained models under two criteria: same-session and cross-session. This analysis employs a one-second window size and explores both user-provided and estimated labeling methods for determining emotion elicitation time.

Initially, user-provided labels are used for the experiments. Table 6.2 presents the results of the 10% proportional fine-tuning with user-provided labels. The multi-representation approach consistently outperforms all baselines on both same-session and cross-session criteria. This method achieves an F1-score of 0.84 for same-session and 0.70 for cross-session, surpassing the periocular-only baseline (0.82 and 0.54, respectively). Notably, while the periocular-only baseline achieves comparable performance to the multi-representation approach under same-session criteria, it is significantly outperformed in the more challenging and practical cross-session criteria. The eye movement-only and content attention baselines demonstrate limited performance under the one-second window size, with cross-session F1-scores of only 0.23 and 0.17, respectively, indicating their inadequacy for practical emotion recognition tasks in this context.

To simulate more practical scenarios where user labels are unavailable during personalized adaptation, a comparison is conducted between user-provided labels and estimated labels. Table 6.3 presents this comparison for the multi-representation method and the periocular-only baseline under 10% proportional fine-tuning. The results demonstrate remarkably similar performance regardless of the labeling method used, suggesting that estimated labels derived from other users' data can effectively substitute user-provided labels in practical applications while maintaining comparable emotion recognition performance.

To further validate the effectiveness of the estimated labeling method, a statistical analysis is conducted. A Wilcoxon Signed-Rank Test comparing the F1-scores of the multi-representation model using both labeling methods delivered a p value of 0.125, showing no statistically significant difference between user-provided (Mean = 0.71, Deviation = 0.12) and estimated (Mean = 0.73, Deviation = 0.12) labels. This analysis supports the conclusion that the estimated labeling method can effectively substi-

Table 6.3: Comparison of user-provided and estimated labeling methods

Labeling	Modalities	10% Same-Session	10% Cross-Session
User	Periocular	0.78	0.52
	Eye Movement + Periocular	0.84	0.70
Estimated	Periocular	0.83	0.54
	Eye Movement + Periocular	0.85	0.71

Table 6.4: Few-shot fine-tuning performance: User-provided vs. Estimated Labels

Labeling	Modalities	1-shot	2-shot	3-shot*	4-shot*	5-shot*
User	Periocular	0.53	0.56	0.57	0.57	0.57
	Eye Movement + Periocular	0.67	0.68	0.69	0.70	0.70
Estimated	Periocular	0.55	0.56	0.57	0.57	0.57
	Eye Movement + Periocular	0.68	0.70	0.70	0.71	0.71

*: For emotion 'Surprise', at most 2 shots are taken.

tute user-provided labels in practical applications.

Moreover, an investigation on the impact of labeling methods in a few-shot fine-tune cross-session setting is also conducted, as shown in Table 6.4. This setting evaluates the model's ability to generalize emotion recognition to unseen content while training on only a small proportion of data from seen content. The multi-representation approach consistently and significantly outperforms the periocular-only baseline across all shot settings. With user-provided labels, the multi-representation method achieves an F1-score of 0.67 under 1-shot, compared to 0.53 for the periocular-only method. This performance gap persists as the number of shots increases, with the multi-representation method reaching an F1-score of 0.70 at the 4-shot setting, while the periocular-only method peaks at 0.57.

Focusing on the more realistic 4-shot and 5-shot settings for practical applications, an overall comparison of both modalities combined shows similar performance between user-provided labels (Mean = 0.635, Standard Deviation = 0.065) and estimated labels (Mean = 0.640, Standard Deviation = 0.070). A Wilcoxon Signed-Rank Test yielded a p-value of 0.157, suggesting no statistically significant difference between the two labeling methods.

These results highlight the proposed multi-representation method's effective generalization capability with very limited training data and its robustness in real-world scenarios where ground truth labels may be unavailable during application.

6.3. Ablation Studies

6.3.1. Impact of Eye Movement Window Size

The choice of window size for eye movement data analysis is a critical factor in emotion recognition tasks. While the proposed method in this study employs short window sizes due to computational constraints imposed by the video feature extractor, it is important to understand the potential impact of longer window sizes on the performance. Previous studies have explored a range of window sizes for emotion recognition tasks, with some extending up to 10 seconds in comparative studies [40] and even 180 seconds (three minutes) in research on emotion elicitation in virtual reality environments [84]. To assess the impact of window size on our model's performance, the eye movement-only baseline is evaluated for various window sizes up to 15 seconds. Table 6.5 presents the results of this analysis, revealing a consistent trend of performance improvement as the window size increases.

In the pre-train stage, a substantial improvement in performance as the window size increases is observed. The F1-score rises from 0.26 at 0.5 seconds to 0.39 at 15 seconds, representing a 50% improvement. This trend suggests that longer window sizes continue to provide additional informative features for emotion recognition, potentially capturing more temporal patterns in eye movements that are indicative of emotional states. The fine-tuning stage exhibits a similar pattern of improvement. For the 10% same-session scenario, the F1-score increases from 0.46 at 1 second to 0.54 at 15 sec-

Table 6.5: Eye movement representation performance: Window size analysis

Modalities	Stage	Window Size (s)						
		0.5	1	1.5	2	5	10	15
Eye Movement	Pre-train	0.26	0.27	0.29	0.30	0.34	0.37	0.39
	Fine-tune*	0.46	0.46	0.44	0.43	0.50	0.52	0.54

*: 10% same-session fine-tune.

onds, a 17.4% increase. This improvement indicates that personalized adaptation benefits from longer temporal contexts, allowing the model to capture individual-specific patterns in eye movements over extended periods. However, it is crucial to note that even with a 15-second window, the absolute performance (0.39 for pre-train, 0.54 for fine-tune) still does not surpass the proposed approach with shorter windows.

6.3.2. Effect of Periocular Recording Frame Rate

The core intuition in this research is that directly feeding high frame rate (120fps) periocular recordings into a model is computationally infeasible. Instead, the approach samples the recordings at a lower frame rate and uses high-frequency eye movement data to compensate for the loss of frames. To test the hypothesis that higher frame rate input video should lead to better performance, an evaluation is conducted on the impact of varying the frame rate of periocular recordings while maintaining eye movement data at 120Hz, with a constant window size of one second. For this experiment, the default model configuration as described in Section 5.6 is maintained for all frame rates except 2fps. In the 2fps setting, the frame patch size is adjusted from 5 to 2 frames temporally. This adjustment is primarily necessary to accommodate the input shape requirements of the model. The modification allows configurations (5fps with 5-frame patches and 2fps with 2-frame patches) to treat a one-second video segment as a single temporal unit. By maintaining this consistent temporal treatment across frame rates, a fair comparison of the model's ability to extract meaningful features from equivalent time spans is ensured.

Table 6.6 presents the results for 10% proportional fine-tuning across different frame rates. The results confirm the hypothesis: higher frame rates generally lead to improved performance for both the periocular-only baseline and the proposed multi-representation method. The proposed approach consistently outperforms the periocular-only approach across all frame rates and fine-tuning scenarios. In the more practically relevant cross-session scenario, the proposed method shows significant improvements over the periocular-only method. At 2fps, the proposed method achieves an F1-score of 0.67, compared to 0.52 for the periocular-only approach, featuring a 28.8% improvement. This performance gap is maintained across higher frame rates, with the proposed method reaching an F1-score of 0.71 at 20fps, compared to 0.56 for the periocular-only method. It's worth noting that the performance gains from increasing frame rates tend to diminish at higher frame rates. For the proposed approach, the improvement from 2fps to 10fps (0.67 to 0.70) is more substantial than from 10fps to 20fps (0.70 to 0.71). This suggests that the proposed method can achieve strong performance even at lower frame rates, which has positive implications for computational efficiency.

6.4. Profiling

To assess the practical viability of the proposed multi-representation approach, a comprehensive profiling analysis is conducted on model variations that demonstrated strong performance in the practical cross-session scenario. The profiling focused on the inference phase, examining the impact of varying periocular recording frame rates while maintaining eye movement data at 120Hz, with a constant window size of one second.

The profiling is performed on an Alienware Desktop equipped with an 11th Gen Intel Core i7-11700KF 3.60GHz CPU and an NVIDIA GeForce RTX 3080Ti GPU with 12GB of memory. This hardware configuration aligns with typical desktop configuration used in combination with the VR headset setups, which ensures the relevance of the results to real-world applications.

Parameters and model sizes are directly measured, as these remain constant across inferences.

Table 6.6: Performance analysis: Varying frame rates for periocular and combined representation

Modalities	Frame Rate	10% Same-session	10% Cross-Session
Periocular	2fps	0.78	0.52
	5fps	0.79	0.52
	10fps	0.82	0.54
	15fps	0.83	0.58
	20fps	0.84	0.56
Eye Movement + Periocular	2fps	0.80	0.67
	5fps	0.80	0.68
	10fps	0.84	0.70
	15fps	0.84	0.70
	20fps	0.84	0.71

Table 6.7: Profiling of proposed multi-representation (peiocular + eye eovements) models at various frame rates

Frame Rate	Parameters	Model Size	Memory Required	Inference Time
2fps	4,271,757	17.02MB	342MB	3.30ms
5fps	4,469,901	17.81MB	342MB	3.39ms
10fps	4,520,077	18.01MB	344MB	3.55ms
15fps	4,570,253	18.21MB	364MB	3.56ms
20fps	4,620,429	18.42MB	366MB	3.61ms

Memory requirements and inference time are averaged over 9,000 iterations, with the first 1,000 iterations discarded to mitigate any latency effects from initial model loading. Inference time is measured using Python's built-in timing functions, while memory consumption is monitored using Nvidia's proprietary application.

Table 6.7 presents the profiling results. The data reveal a clear trend: as the frame rate increases, all measured factors show an increase. It is noteworthy that even at the highest frame rate (20fps), the model maintains performance well within the bounds required for real-time emotion recognition. In other words, the inference time remains far below 50ms, enabling the system to perform a new emotion recognition for each incoming frame.

7 Discussion

7.1. Discussion on Results

The results presented above in Chapter 6 reveal two clear trends. Firstly, the proposed multi-representation approach consistently outperforms all other baselines. Secondly, by utilizing more information, specifically longer window sizes and higher frame rates of periocular records, leads to improved performance. However, several phenomena are observed, which includes the slight out-performance of other baselines under the pre-train settings, the consistency of the estimated labeling method out-performing the user-provided labeling method, and the saturation threshold of factors such as window size, recording frame rate, and frame size on the performance.

7.1.1. Pre-train Settings and Personal Features

The intuition to the phenomenon observed in the pre-train settings is that emotion recognition is highly dependent on personal features. All methods perform poorly in pre-train settings, with the highest accuracy reaching only 0.52 using the proposed multi-representation approach with a two-second window size. However, further fine-tuning with a single sample from each emotion class significantly boosts performance. For instance, the proposed method and the method utilizing only periocular recordings both achieve 0.44 accuracy under pre-train settings with 10fps recordings, 120Hz eye movements, and a one-second window. In contrast, under the fine-tune setting of cross-session 1-shot, their performance increases to 0.67 and 0.53, respectively. This significant improvement from a single sample, representing only one second of data from the subject, underscores the importance of personal features in emotion recognition tasks and the necessity of fine-tuning for accurate performance evaluation.

7.1.2. Estimated vs. User-provided Labeling Methods

The consistent out-performance of the estimated labeling method is likely due to that it does not contain extreme situations where subjects report no emotion elicited in an emotion session. The estimated labeling method ensures that emotions are attributed to all sessions, while using user-provided labels include sessions that are reported as no emotion elicited. This causes troubles in the training phase in cross-session scenarios and leading to lower overall performance.

7.1.3. Saturation Thresholds and Resource Constraints

Some evidence, as presented in Table 6.6, suggests that periocular recording frame rates above 10fps yield minimal performance improvements. However, this aspect remains incompletely explored, together with other factors such as window sizes. The primary constraint is the memory-intensive nature during training of models that utilize video information. The limit of the training system is under the setting of 20fps periocular recordings, 120Hz eye movements with a one-second window size.

7.1.4. Comparison with Existing Studies

To facilitate a comparison with existing studies [30, 84, 96], the performance of valence and intensity identification is performed. The proposed approach, integrating periocular recordings at 10fps and eye

movements at 120Hz, achieves a performance of 0.70 in the 5-shot cross-session scenario with user-provided labels. After categorizing recognized emotions on valence and intensity scales, the approach demonstrates performance of 0.89 for valence identification and 0.81 for intensity identification. These results are comparable to those reported in other studies, despite the significantly shorter window size of one second used in this study compared to the ten-second to minute-long windows employed in others.

7.2. Discussion on Model Decision-Making

To explore the explanability of the decision-making of proposed multi-representation approach, further analysis are conducted. This section aims to discuss the reason why the multi-representation approach significantly outperforms the baseline method that utilizes only periocular recordings (referred to as the single-representation approach in this section).

For these analysis, data from the first subject (Subject ID P002) is used with user-provided labels. The more practical cross-session scenario is chosen, specifically employing 5-shot training on one session and testing with data from another session of the same emotion. The analysis are performed on data with a 10fps frame rate for periocular recordings while maintaining eye movement data at 120Hz, with a constant window size of one second.

7.2.1. t-SNE Analysis

t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis is first performed to visualize the high-dimensional data in a two-dimensional space. This technique is particularly useful in this case as it helps to reveal patterns in the data by clustering similar points together while maintaining the relative distances between dissimilar points. The features for t-SNE analysis are extracted from the output of the module immediately before the passing into the emotion classifier, as previously illustrated in Figure 5.5.

Figure 7.1 presents the results of the t-SNE analysis. Figure 7.1(a) shows the features extracted by the proposed multi-representation approach, while Figure 7.1(b) and (c) display those extracted by the single-representation baselines. The multi-representation approach demonstrates a clearer division among features extracted from each emotion. Both multi-representation and periocular approaches struggle to extract distinct features for neutral and surprise emotions, but the multi-representation approach shows clear divisions among the other emotions, whereas the periocular approach fails to establish distinct boundaries between features extracted among disgust, anger, and sadness. The approach using only eye movements fails to draw any clear divisions among the features extracted.

This observation suggests that while periocular recordings alone perform well, the additional information contained in eye movements provides crucial cues for distinguishing between certain emotions, particularly disgust, fear, and anger.

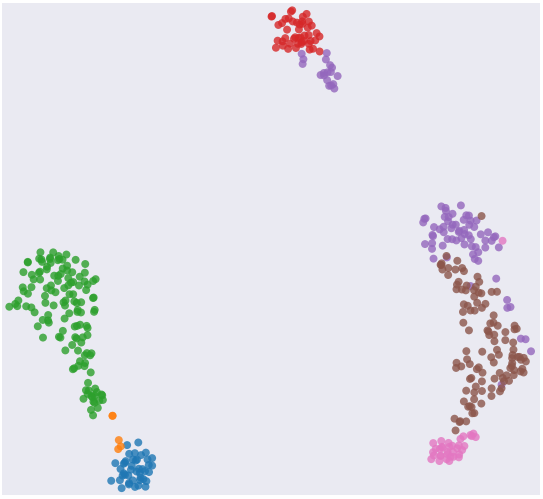
7.2.2. Attention Analysis

Attention analysis is also conducted to further understand the model's focus during emotion recognition. As defined in Section 5.6, the temporal size for patches is five, resulting in two tokens extracted from each temporal dimension that represent the collective information of a set of frames. These tokens are utilized to plot the attention heat-map. Three versions of the heat-map were created: two using each of the individual spatial tokens, and one using the mean of both tokens. Detailed information on the realization of this visualization can be found in Appendix A.6.

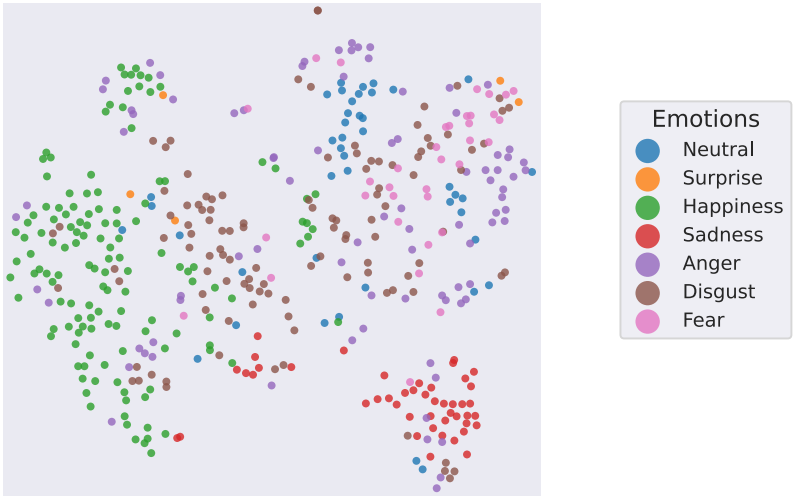
Figure 7.2 presents the visualization of the attention heat-maps, where white indicates high attention and black indicates low attention. The multi-representation approach consistently focuses on the areas surrounding the eyes across all emotions. In contrast, the single-representation approach does not exhibit a clear attention pattern. This difference is particularly evident in the happiness and sadness emotions.

These findings suggest that the information in eye movements may enable the model to consistently focus on the areas surrounding the eyes, as the eye movements already contain informative information of the eyes. This consistent attention potentially contributes to the improved performance of the multi-representation approach.

(a) Multi-representation t-SNE Analysis



(a) Periocular-representation t-SNE Analysis



(c) Eye-movement-representation t-SNE Analysis

Figure 7.1: t-SNE analysis visualization of emotion features extracted from (a) multi-representation approach, (b) single-representation approach with periocular, and (c) single-representation approach with eye movements. The analysis is performed on data from Subject P002 in a five-shot cross-session scenario. Features are extracted from the output of the fused feature module before emotion classification.

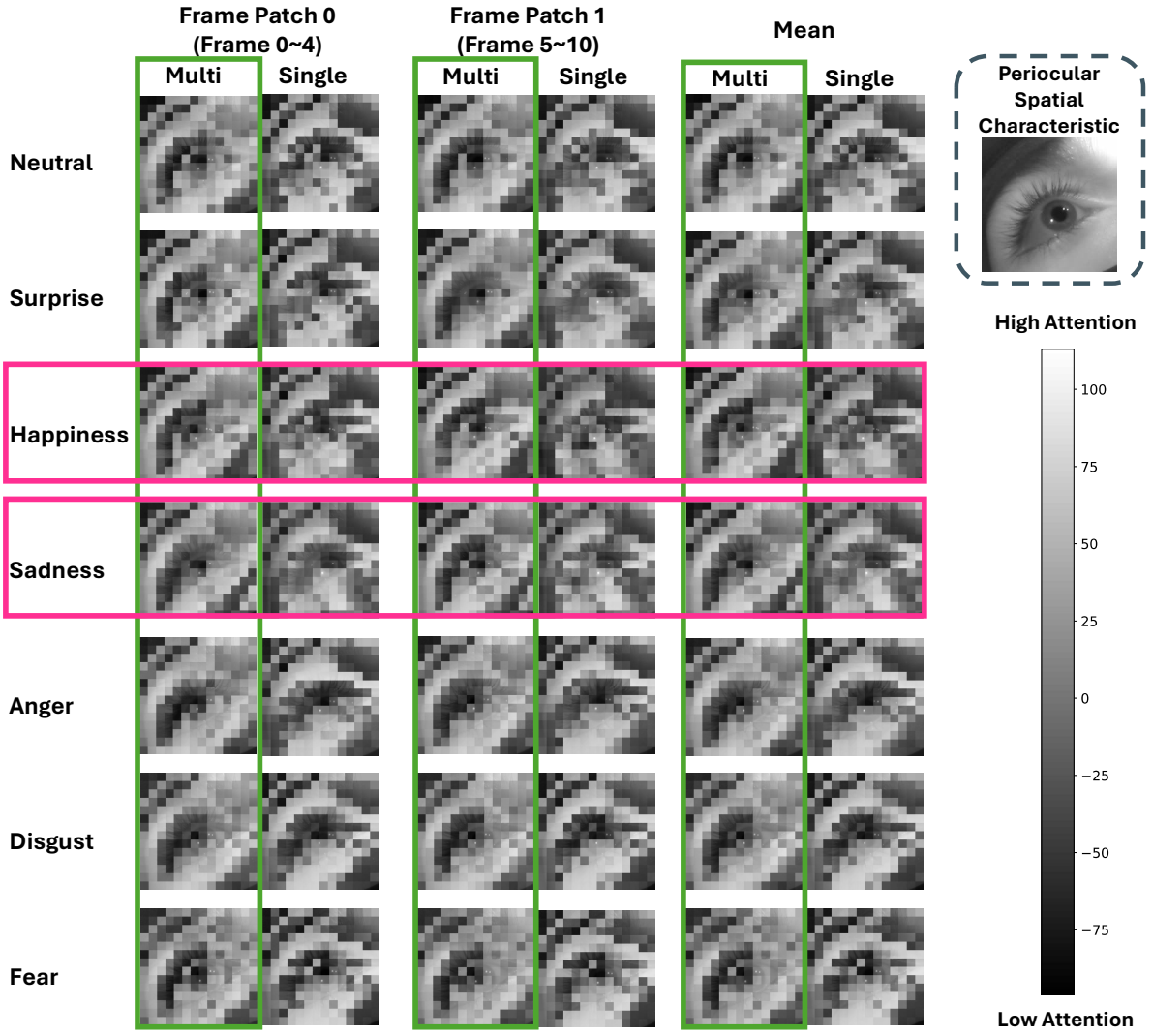


Figure 7.2: Attention visualization comparison between multi-representation and single-representation approaches for Subject P002 across all seven emotions, focusing on the right eye. The analysis uses a one-second window size, with periocular recordings sampled at 10fps and eye movements at 120Hz in a five-shot cross-session experiment setup.

7.3. Limitations

This study acknowledges two primary limitations: the lack of comparable methods and the absence of in-the-wild experiments.

7.3.1. Lack of Comparable Methods

The comparative analysis in this study is limited due to the evaluation solely being performed on the dataset specifically gathered for it. Moreover, the methods utilizing different representations are all variants of the proposed approach, which restricts the range of comparison. This limitation is primarily due to the fact of the absence of publicly available datasets that offer the same data representations for emotion recognition in immersive environments. The lack of comparability with other studies in the field potentially limits the ability to evaluate the performance achieved in this study with other research. This limitation may reduce the confidence in claiming how the current results compare to existing benchmarks in the field.

7.3.2. Absence of In-the-Wild Experiments

The current study does not include experiments conducted in naturalistic settings, such as participants watching videos or performing daily tasks in a VR environment. While it is common practice to gather data and conduct experiments under controlled and standardized conditions, this approach may be influenced by various factors, including the careful selection of stimuli presented to subjects, specific hardware configurations, and the installation of eye-tracking devices. The controlled environment, while providing standardization, may not fully represent real-world scenarios.

8

Conclusion & Recommendation

8.1. Conclusion

This study proposes an approach for emotion recognition in immersive environments that utilizes solely multi-representation data extracted from users' eyes. By combining periocular recordings from both eyes with eye movement data, the method effectively addresses the challenge of achieving promising results in recognizing seven distinct emotions.

The proposed multi-representation method leverages the complementary information from eye movements and periocular recordings, significantly outperforming baselines that use only one type of data or incorporate content stimuli. Notably, in the same-session scenario, where the model is tested on responses to unseen parts of the stimuli used for adaptation, the approach achieves an F1-score of 0.85 using only 10% of the data for personal adaptation. Furthermore, to enhance user experience and minimize the amount of data required from each user, it has been demonstrated that the proposed method achieves robust performance with minimal adaptation data. Using only five seconds of data from each emotion for personal adaptation, summing to 35 seconds in total. The proposed method achieves an F1-score of 0.71 in cross-session emotion recognition, where the model is tested on responses to completely different stimuli from the one used for adaptation. Under the same but more extreme condition, where only one second of data is available from each emotion, seven seconds in total, the proposed method achieves an F1-score of 0.68.This indicates that the approach can quickly adapt to new users and perform effectively on unseen content with very limited data.

Importantly, the conducted experiments also show that using estimated labels derived from other users' data can effectively substitute for user-provided labels in the personal adaption process. Statistical analysis confirms that there is no significant difference in performance between using user-provided labels and estimated labels, highlighting the practicality of the method in real-world applications where obtaining user labels may not be feasible.

In conclusion, this study demonstrates the effectiveness of employing multi-representation behavioral responses from users' eyes for emotion recognition in immersive environments. By requiring minimal data for personal adaption, the proposed approach offers a practical and efficient solution for real-world applications. The promising results underscore the potential of the method to enhance user experience and interaction in immersive technologies.

8.2. Future Work

Future work can be performed in three main directions: enhancing model performance, investigating model explainability, and developing real-time emotion recognition systems for in-the-wild experiments.

To build upon the baseline set by the proposed multi-representation approach, further research can explore the allocation of additional computational resources to train larger models and evaluate their performance. Moreover, investigations on architectural modifications can be performed to enhance both the feature extraction process and the fusion of features from different representations.

A deeper examination of model explainability is another direction for future work. It can provide valuable insights into the decision-making processes of emotion recognition. This involves identifying and extracting critical features that contribute to the recognition of specific emotions. Furthermore,

methods for weighting and visualizing the precise contributions of each data representation to the final recognized emotion can also be developed. These efforts would not only enhance the interpretability of the model but also potentially improving the design for more effective emotion recognition systems.

The development of a real-time emotion recognition system for in-the-wild experiments is another direction for further research. It can not only validate the performance reported in the current study through practical, real-world experiments, but also potentially gather a new dataset from users performing a wider range of common tasks in immersive environments. This requires a design of a new labeling technique that demands less user effort to adapt the nature of other tasks performed in immersive environments. The newly gathered dataset can provide a richer and more diverse set of data for future emotion recognition systems to be evaluated on.

8.3. Implications for HCI and Immersive Environments

The promising performance achieved in emotion recognition could lead to developing more engaging, more adaptive and more personalized experience in immersive environments.

Accurate emotion recognition could benefit enhancing the sense of presence and improving social interactions in immersive environments. Research has shown that the realistic appearance of avatars, including vivid facial expressions, enhances the self-perception of one's own virtual body and leads to more positive communication experiences in virtual spaces [45]. Moreover, a positive correlation between highly embodied avatars and users' deep engagement in immersive environments has also be discovered [8]. Promising emotion recognition could contribute to enhancing more realistic facial expressions leading to more embodied avatars. This could improve the current existing social gatherings, remote collaborative work experience in immersive environments.

The ability to accurately recognize emotions in immersive environments also enables the development of adaptive and personalized applications. For instance, emotion logging systems could be implemented for mental state tracking, offering tools for well-being monitoring and clinical purposes. As for the gaming industry, immersive games could dynamically adjust their difficulty, narrative, or environmental elements depending on the emotional state of users at different time points.

While the potential benefits are significant, ethical considerations are also raised. Concerns considering privacy related to the collection and use of personal emotional data should be properly taken care of. Moreover, the potential lead to manipulation in emotion-aware applications for clinical purposes or well-being monitoring should also be thoroughly examined before applying in practice.

References

[1] Mojtaba Khomami Abadi et al. “DECAF: MEG-based multimodal database for decoding affective physiological responses”. In: *IEEE Transactions on Affective Computing* 6.3 (2015), pp. 209–222.

[2] Hamdi Ben Abdessalem et al. “Toward real-time system adaptation using excitement detection from eye tracking”. In: *Intelligent Tutoring Systems: 15th International Conference, ITS 2019, Kingston, Jamaica, June 3–7, 2019, Proceedings 15*. Springer. 2019, pp. 214–223.

[3] Anurag Arnab et al. “Vivit: A video vision transformer”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6836–6846.

[4] Janice Attard-Johnson, Caoilte Ó Ciardha, and Markus Bindemann. “Comparing methods for the analysis of pupillary response”. In: *Behavior Research Methods* 51 (2019), pp. 83–95.

[5] Anthony Bagnall et al. “The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances”. In: *Data mining and knowledge discovery* 31 (2017), pp. 606–660.

[6] Emad Barsoum et al. “Training deep networks for facial expression recognition with crowd-sourced label distribution”. In: *Proceedings of the 18th ACM international conference on multimodal interaction*. 2016, pp. 279–283.

[7] Donald J Berndt and James Clifford. “Using dynamic time warping to find patterns in time series”. In: *Proceedings of the 3rd international conference on knowledge discovery and data mining*. 1994, pp. 359–370.

[8] Frank Biocca. “Connected to My Avatar: Effects of avatar embodiments on user cognitions, behaviors, and self construal”. In: *Social Computing and Social Media: 6th International Conference, SCSM 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings 6*. Springer. 2014, pp. 421–429.

[9] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[10] Margaret M. Bradley and Peter J. Lang. “Measuring emotion: The self-assessment manikin and the semantic differential”. In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59. ISSN: 0005-7916.

[11] Sven Buechel and Udo Hahn. “Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis”. In: *arXiv preprint arXiv:2205.01996* (2022).

[12] Felix Burkhardt et al. “A database of German emotional speech.” In: *Interspeech*. Vol. 5. 2005, pp. 1517–1520.

[13] Carlos Busso et al. “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language resources and evaluation* 42 (2008), pp. 335–359.

[14] Shiyang Cheng et al. “4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5117–5126.

[15] Roddy Cowie et al. “’FEELTRACE’: An instrument for recording perceived emotion in real time”. In: *Proc. ITRW on Speech and Emotion*. 2000, pp. 19–24.

[16] Nele AJ De Witte et al. “Augmenting exposure therapy: Mobile augmented reality for specific phobia”. In: *Frontiers in Virtual Reality* 1 (2020), p. 8.

[17] Rami Reddy Devaram et al. “LEMON: a lightweight facial emotion recognition system for assistive robotics based on dilated residual convolutional neural networks”. In: *Sensors* 22.9 (2022), p. 3366.

[18] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019.

[19] Julia Diemer et al. “The impact of perception and presence on emotional reactions: a review of research in virtual reality”. In: *Frontiers in psychology* 6 (2015), p. 26.

[20] Miranda R Donnelly et al. “Virtual reality for the treatment of anxiety disorders: a scoping review”. In: *The American Journal of Occupational Therapy* 75.6 (2021).

[21] Alexey DOSOVITSKIY. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[22] Samira Ebrahimi Kahou et al. “Recurrent neural networks for emotion recognition in video”. In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015, pp. 467–474.

[23] Paul Ekman. “An argument for basic emotions”. In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[24] Paul Ekman and Wallace Friesen. “Constants across cultures in the face and emotion”. In: *Journal of personality and social psychology* 17 (Feb. 1971), pp. 124–9.

[25] Moomal Farhad et al. “A review of medical diagnostic video analysis using deep learning techniques”. In: *Applied Sciences* 13.11 (2023), p. 6582.

[26] Carlos Flavián, Sergio Ibáñez-Sánchez, and Carlos Orús. “The impact of virtual, augmented and mixed reality technologies on the customer experience”. In: *Journal of business research* 100 (2019), pp. 547–560.

[27] Sarvenaz Ghafourian, Ramin Sharifi, and Amirali Baniasadi. “Facial emotion recognition in imbalanced datasets”. In: *Computer Science and Information Technology* (2022).

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[29] Quentin Guimard et al. “Pem360: A dataset of 360 videos with continuous physiological measurements, subjective emotional ratings and motion traces”. In: *Proceedings of the 13th ACM Multimedia Systems Conference*. 2022, pp. 252–258.

[30] Kunal Gupta et al. “Total vrecall: Using biosignals to recognize emotional autobiographical memory in virtual reality”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (2022), pp. 1–21.

[31] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[32] Ruining He et al. *RealFormer: Transformer Likes Residual Attention*. 2021.

[33] Hassan Ismail Fawaz et al. “Deep learning for time series classification: a review”. In: *Data mining and knowledge discovery* 33.4 (2019), pp. 917–963.

[34] Carroll E Izard. “Emotion theory and research: Highlights, unanswered questions, and emerging issues”. In: *Annual review of psychology* 60.1 (2009), pp. 1–25.

[35] Carroll E. Izard. *Human Emotions*. New York: Springer US., 1977.

[36] Samira Ebrahimi Kahou et al. “Combining modality specific deep neural networks for emotion recognition in video”. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 543–550.

[37] Fazle Karim et al. “LSTM fully convolutional networks for time series classification”. In: *IEEE access* 6 (2017), pp. 1662–1669.

[38] Andrej Karpathy et al. “Large-scale video classification with convolutional neural networks”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[39] Moritz Kassner, William Patera, and Andreas Bulling. “Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction”. In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 2014, pp. 1151–1160.

[40] Panayu Keelawat et al. “A comparative study of window size and channel arrangement on EEG-emotion recognition using deep CNN”. In: *Sensors* 21.5 (2021), p. 1678.

[41] Sander Koelstra et al. “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[42] Mariska E Kret and Elio E Sjak-Shie. “Preprocessing pupil size data: Guidelines and code”. In: *Behavior research methods* 51 (2019), pp. 1336–1342.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[44] Ivan Laptev. “On space-time interest points”. In: *International journal of computer vision* 64 (2005), pp. 107–123.

[45] Marc Erich Latoschik et al. “The effect of avatar realism in immersive social virtual realities”. In: *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. VRST ’17. Gothenburg, Sweden: Association for Computing Machinery, 2017. ISBN: 9781450355483.

[46] Raymond Lavoie et al. “Virtual experience, real consequences: the potential negative emotional consequences of virtual reality gameplay”. In: *Virtual Reality* 25.1 (2021), pp. 69–81.

[47] Séverin Lemaignan et al. “Artificial cognition for social human–robot interaction: An implementation”. In: *Artificial Intelligence* 247 (2017), pp. 45–69.

[48] Christophe Leys et al. “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median”. In: *Journal of Experimental Social Psychology* 49.4 (2013), pp. 764–766. ISSN: 0022-1031.

[49] Chao Li et al. “Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 1240–1244.

[50] Yihao Li et al. “A review of deep learning-based information fusion techniques for multimodal medical image classification”. In: *Computers in Biology and Medicine* (2024), p. 108635.

[51] Wei Liu et al. “Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition”. In: *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[52] Yuanyuan Liu et al. “Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 24–32.

[53] Tuva Fjærtoft Lønne et al. “The effect of immersion on sense of presence and affect when experiencing an educational scenario in virtual reality: A randomized controlled study”. In: *Heliyon* 9.6 (2023).

[54] Hui Ma et al. “A transformer-based model with self-distillation for multimodal emotion recognition in conversations”. In: *IEEE Transactions on Multimedia* (2023).

[55] Jiaxin Ma et al. “Emotion recognition using multimodal residual LSTM network”. In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 176–183.

[56] Fazliddin Makhmudov, Alpamis Kultimuratov, and Young-Im Cho. “Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures”. In: *Applied Sciences* 14.10 (2024), p. 4199.

[57] Ibrahim Malik et al. “Emotions beyond words: Non-speech audio emotion recognition with edge computing”. In: *arXiv preprint arXiv:2305.00725* (2023).

[58] Sebastiaan Mathôt et al. “Safe and sensible preprocessing and baseline correction of pupil-size data”. In: *Behavior research methods* 50 (2018), pp. 94–106.

[59] Anjela Mayer et al. “Collaborative work enabled by immersive environments”. In: *New Digital Work: Digital Sovereignty at the Workplace*. Springer International Publishing Cham, 2023, pp. 87–117.

[60] Albert Mehrabian and James A. Russell. “The Basic Emotional Impact of Environments”. In: *Perceptual and Motor Skills* 38.1 (1974). PMID: 4815507, pp. 283–301.

[61] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. “Affectnet: A database for facial expression, valence, and arousal computing in the wild”. In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[62] Seyed Muhammad Hossein Mousavi et al. “Emotion Recognition in Adaptive Virtual Reality Settings: Challenges and Opportunities.” In: *WAMWB@ MobileHCI* (2023), pp. 1–20.

[63] Peter Mundy et al. “Defining the social deficits of autism: The contribution of non-verbal communication measures”. In: *Journal of child psychology and psychiatry* 27.5 (1986), pp. 657–669.

[64] Alex Nanopoulos, Rob Alcock, and Yannis Manolopoulos. “Feature-based classification of time-series data”. In: *International Journal of Computer Research* 10.3 (2001), pp. 49–61.

[65] Anneli Olsen. “The Tobii IVT Fixation Filter Algorithm description”. In: *2012 Tobii Technology*. 2012.

[66] Francisco Javier Ordóñez and Daniel Roggen. “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition”. In: *Sensors* 16.1 (2016), p. 115.

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

[68] Vlad Pandelea et al. “Emotion recognition on edge devices: Training and deployment”. In: *Sensors* 21.13 (2021), p. 4496.

[69] Alexander M Pascual et al. “Light-FER: a lightweight facial emotion recognition system on edge devices”. In: *Sensors* 22.23 (2022), p. 9524.

[70] Katarina Pavic et al. “Feeling virtually present makes me happier: The influence of immersion, sense of presence, and video contents on positive emotion induction”. In: *Cyberpsychology, Behavior, and Social Networking* 26.4 (2023), pp. 238–245.

[71] Soujanya Poria et al. “Meld: A multimodal multi-party dataset for emotion recognition in conversations”. In: *arXiv preprint arXiv:1810.02508* (2018).

[72] Hugo Proença and João C. Neves. “Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks”. In: *IEEE Transactions on Information Forensics and Security* 13.4 (2018), pp. 888–896.

[73] Gabriela Maria Pyjas, Jonathan Weinel, and Martyn Broadhead. “Storytelling and VR: Inducing emotions through AI characters”. In: *Proceedings of EVA London 2022*. BCS Learning & Development. 2022, pp. 198–204.

[74] Bishwas Regmi. “Emotion Recognition in Virtual Reality”. MA thesis. TUDelft, 2024.

[75] Oliver Richter and Roger Wattenhofer. *Normalized Attention Without Probability Cage*. 2020.

[76] James A Russell. “A circumplex model of affect”. In: *Journal of Personality and Social Psychology* 39.6 (1980), p. 1161.

[77] Asmaa Sakr and Tariq Abdullah. “Virtual, augmented reality and learning analytics impact on learners, and educators: A systematic review”. In: *Education and Information Technologies* (2024), pp. 1–50.

[78] Patrick Schäfer. “The BOSS is concerned with time series classification in the presence of noise”. In: *Data Mining and Knowledge Discovery* 29 (2015), pp. 1505–1530.

[79] Caspar M Schwiedrzik and Sandrin S Sudmann. “Pupil diameter tracks statistical structure in the environment to increase visual sensitivity”. In: *Journal of Neuroscience* 40.23 (2020), pp. 4565–4575.

[80] Karen Simonyan and Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. In: *Advances in neural information processing systems* 27 (2014).

[81] Henrique Siqueira, Sven Magg, and Stefan Wermter. “Efficient facial feature learning with wide ensemble-based convolutional neural networks”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 5800–5809.

[82] Stuart R Steinhauer et al. “Publication guidelines and recommendations for pupillary measurement in psychophysiological studies”. In: *Psychophysiology* 59.4 (2022), e14035.

[83] Ekaterina Sviridova et al. “Immersive technologies as an innovative tool to increase academic success and motivation in higher education”. In: *Frontiers in Education*. Vol. 8. Frontiers Media SA. 2023, p. 1192760.

[84] 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).

[85] Luma Tabbaa et al. “Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures”. In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5.4 (2021), pp. 1–20.

[86] Ying Tan et al. “A multimodal emotion recognition method based on facial expressions and electroencephalography”. In: *Biomedical Signal Processing and Control* 70 (2021), p. 103029.

[87] Panagiotis Tzirakis et al. “End-to-end multimodal emotion recognition using deep neural networks”. In: *IEEE Journal of selected topics in signal processing* 11.8 (2017), pp. 1301–1309.

[88] DWF Van Krevelen and Ronald Poelman. “A survey of augmented reality technologies, applications and limitations”. In: *International journal of virtual reality* 9.2 (2010), pp. 1–20.

[89] A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).

[90] Subhashini Venugopalan et al. “Sequence to sequence-video to text”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4534–4542.

[91] Dongrui Wu et al. “Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality stroop task”. In: *IEEE Transactions on Affective Computing* 1.2 (2010), pp. 109–118.

[92] Ruibin Xiong et al. *On Layer Normalization in the Transformer Architecture*. 2020.

[93] Huijuan Xu, Abir Das, and Kate Saenko. “R-c3d: Region convolutional 3d network for temporal activity detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5783–5792.

[94] Jieying Xue et al. “BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks”. In: *arXiv preprint arXiv:2407.04279* (2024).

[95] Tong Xue et al. “CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360VR Videos”. In: *IEEE Transactions on Multimedia* 25 (2021), pp. 243–255.

[96] Tong Xue et al. “Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels”. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–15.

[97] Taeyang Yun et al. “TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation”. In: *arXiv preprint arXiv:2401.12987* (2024).

[98] George Zerveas et al. *A Transformer-based Framework for Multivariate Time Series Representation Learning*. 2020.

[99] George Zerveas et al. “A transformer-based framework for multivariate time series representation learning”. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2114–2124.

[100] Haiwei Zhang et al. “In the blink of an eye: Event-based emotion recognition”. In: *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, pp. 1–11.

[101] Tianyi Zhang et al. “Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–15.

[102] Wei Zhang et al. “Transformer-based multimodal information fusion for facial expression analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2428–2437.

[103] Jianfeng Zhao, Xia Mao, and Lijiang Chen. “Speech emotion recognition using deep 1D & 2D CNN LSTM networks”. In: *Biomedical signal processing and control* 47 (2019), pp. 312–323.

[104] Wei-Long Zheng et al. “Emotionmeter: A multimodal framework for recognizing human emotions”. In: *IEEE transactions on cybernetics* 49.3 (2018), pp. 1110–1122.

[105] Peixiang Zhong, Di Wang, and Chunyan Miao. “EEG-based emotion recognition using regularized graph neural networks”. In: *IEEE Transactions on Affective Computing* 13.3 (2020), pp. 1290–1301.

A

Appendix

A.1. Video Context Instructions

The following context sentences were read to the participants before each video stimulus to provide background information and prime them for the intended emotional response (excluding Neutral and Surprise), without explicitly mentioning the target emotion:

- 0a Neutral: A man was selected to assist another man in his research, and the two men are discussing the topic on AI.
- 1a Surprise: A woman is riding a bike among quiet and calm streets to meet her husband.
- 1b Surprise: A man is giving a speech trying to unite his fellow scientists to not give up when they find themselves trapped in a research lab in the middle of the ocean.
- 0b Neutral: A group of coworkers are having a business meeting to try and secure a partnership with a fellow company.
- 2a Happiness: A disabled girl is coach surfing by a beach at a rural place in Thailand. She tries to encourage local kids to join her in this amazing sport that gave her hope for life.
- 2b Happiness: A teen bought a lottery ticket for his grandma, and the clerk convinced him to buy one for himself as well.
- 3a Sadness: A lady diagnosed with Alzheimer's is looking at the family album. She can remember her mother and her sister, but not the location of the bathroom in her own house.
- 3b Sadness: A conversation takes place among a girl diagnosed with leukemia, her family, and the doctor. The doctor steps away from the child and informs the mother that her daughter is dying.
- 4a Anger: During the age of colonization, the colonizers treated slaves as animals that could be traded. While the officers are bargaining, a mother of two kids is begging them to save her children, but the officers ignore her and treat her violently.
- 4b Anger: A man arrives home to his wife, who is cooking dinner. She discovers evidence of him cheating and he acts violently in response.
- 5a Disgust: A man with a Nazi tattoo treats a person of color violently and shows no regret. He smiles as the police arrive and arrest him.
- 5b Disgust: Two men have just fought and are lying on the ground face to face. Blood floods the floor, flowing from one man to the other.
- 6a Fear: Three friends built a cabin in the woods. Two of them left for home, while one insists on staying and living there alone.
- 6b Fear: A lady living alone in a house suddenly hears noises coming from the basement and decides to go down to investigate.

A.2. Subjects Emotion Eliciting Time Across Sessions

Table A.2 presents the emotion eliciting time across different sessions for subjects in the study. It provides a detailed breakdown of when subjects reported experiencing emotional responses during each session. The table is organized by session number, time stamp, and subject IDs. This information is derived from user-provided labels.

Table A.1: Subjects emotion eliciting time across sessions

Session No.	Time Stamp	Subject IDs
0a	0:00	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
0b	0:00	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
1a	0:35	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
1b	1:26	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
2a	0:19	2, 3, 6, 8, 11, 12, 14, 15, 18, 24, 26, 28, 32
	1:20	7, 9, 13, 17, 23, 25, 30, 31, 33
2b	0:06	2, 3, 6, 8, 9, 12, 13, 14, 15, 18, 23, 24, 25, 28, 30, 31
	0:45	7, 26
	1:30	11, 33
3a	1:37	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 17, 18, 23, 24, 25, 26, 31, 32, 33
	2:00	15, 28, 30
3b	0:55	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
4a	0:32	2, 6, 7, 8, 9, 11, 12, 13, 14, 15, 18, 23, 25, 26, 28, 30, 31, 32, 33
	1:35	3, 17, 24
4b	0:31	13
	0:56	2, 3, 6, 7, 8, 9, 12, 15, 18, 23, 25, 28, 31, 32, 33
	1:23	11, 14, 24, 26, 30
5a	0:38	2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
5b	0:35	2, 6, 7, 8, 9, 11, 12, 13, 14, 18, 23, 24, 25, 26, 28, 30, 31, 32, 33
6a	0:10	7, 8, 12, 13, 14, 15, 18, 23, 24, 25, 26, 28, 31, 33
	0:28	2, 3, 9
	0:42	11
6b	0:10	2, 7, 8, 9, 11, 14, 15, 17, 18, 23, 24, 26, 28, 30, 31, 32, 33
	0:40	12, 25
	0:47	6, 13

A.3. Pseudo-code for Data Segmentation and Synchronization in Dataset Preparation

This appendix section presents the pseudo-code for data segmentation and synchronization in the dataset preparation process. The piece of code loads various data types including eye videos, gaze data, and pupil data. It then processes these data streams, synchronizing them based on the emotion elicitation time provided by participants. It employs a back-to-front cropping approach to maximize the inclusion of relevant data after the reported elicitation point, and allows for flexible handling of different data types and rates.

```
def segment_and_synchronize_data(subject, session, elicitation_time, window_size):
    # Load data
    # eye0 represents left eye, eye1 represents right eye
    eye0_video = load_video(subject, session, 'eye0')
    eye1_video = load_video(subject, session, 'eye1')
    gaze_data = load_gaze_data(subject, session)
    pupil_data = load_pupil_data(subject, session)

    # Initialize parameters
    fps = 120
    gaze_pupil_rate = 240

    # Calculate end frame and elicitation frame
    end_frame = get_total_frames(eye0_video)
    elicitation_frame = calculate_elicitation_frame(elicitation_time, fps)

    # Crop and process data from back to front
    # This approach maximizes inclusion of relevant data after the user-reported elicitation
        point and mitigates the risk of losing samples that don't fit within the specified
        window size
    eye0_samples = crop_data_back_to_front(eye0_video, end_frame, elicitation_frame,
        window_size, fps)
    eye1_samples = crop_data_back_to_front(eye1_video, end_frame, elicitation_frame,
        window_size, fps, flip='horizontal')
    gaze_samples = crop_data_back_to_front(gaze_data, end_frame, elicitation_frame,
        window_size, gaze_pupil_rate)
    pupil_samples = crop_data_back_to_front(pupil_data, end_frame, elicitation_frame,
        window_size, gaze_pupil_rate)

    # Synchronize samples
    min_samples = min(len(eye0_samples), len(eye1_samples), len(gaze_samples), len(
        pupil_samples))

    synchronized_data = {
        'eye0': eye0_samples[:min_samples],
        'eye1': eye1_samples[:min_samples],
        'gaze': gaze_samples[:min_samples],
        'pupil': pupil_samples[:min_samples]
    }

    return synchronized_data

def crop_data_back_to_front(data, end_point, elicitation_point, window_size, rate):
    samples = []
    current_point = end_point

    while current_point > elicitation_point:
        # The final sample may extend beyond the emotion elicitation time
        # It remains valid as long as it contains data claimed to elicit the emotion
        start_point = max(elicitation_point, current_point - window_size * rate)
        window = extract_window(data, start_point, current_point)

        if isinstance(data, Video) and data.eye == 'eye1':
            window = apply_horizontal_flip(window)

        samples.append(window)
        current_frame -= window_size * rate

    return reverse(samples)  # Reverse to maintain chronological order
```

A.4. Table representation of the output JSON file

Table A.2 presents a tabular representation of the output JSON file generated during the dataset preparation pipeline described in Section 4.3. This JSON file is crucial for efficient data retrieval from the HDF5 files that store individual subject data. The table displays the following information for each participant:

- Participant ID
- Minimum and maximum indices
- Number of samples for each emotion class across two sessions (a and b)

The emotion classes are represented by numbers 0-6, corresponding to the following label map:

- 0: neutral
- 1: surprise
- 2: happiness
- 3: sadness
- 4: anger
- 5: disgust
- 6: fear

This structured representation facilitates a two-step data retrieval process: first, identifying the subject an index belongs to by comparing it to the min and max indices, and second, accessing the data from the corresponding HDF5 file using an adjusted index (calculated as the given index minus the subject's minimum index). The session for each sample can be determined using a similar method.

Table A.2: Subjects sample index of all sessions

Subject	Min Index	Max Index	Emotion #Samples (Session(a)/(b))						
			0	1	2	3	4	5	6
P002	0	869	43/31	5/10	133/108	46/53	85/64	112/13	35/132
P003	870	1514	43/31	4/9	133/107	46/52	17/60	109/0	34/0
P006	1515	2303	43/31	5/10	133/107	45/52	82/62	112/12	0/95
P007	2304	3015	43/31	4/10	73/69	47/53	85/64	112/13	53/55
P008	3016	3901	43/31	4/10	133/108	46/53	84/64	112/13	53/132
P009	3902	4708	43/31	4/10	72/108	46/53	84/64	112/13	35/132
P011	4709	5451	43/31	5/10	133/24	46/52	84/37	112/13	21/132
P012	5452	6311	43/31	4/10	134/108	47/53	85/64	112/13	53/103
P013	6312	7121	43/31	5/10	72/108	46/52	84/89	111/12	53/94
P014	7122	7984	43/31	4/10	133/108	47/53	85/37	113/13	53/133
P015	7985	8834	43/31	5/10	133/108	23/53	84/63	112/0	53/132
P018	8835	9720	43/31	4/9	133/108	47/53	84/64	112/13	53/132
P023	9721	10540	44/31	4/10	72/107	46/53	84/64	109/13	53/130
P024	10541	11306	43/31	4/9	132/105	43/51	17/34	109/10	51/127
P025	11307	12101	43/31	5/10	72/108	45/53	84/64	112/13	53/102
P026	12102	12881	43/30	4/8	128/67	40/48	82/34	105/11	51/129
P028	12882	13740	43/31	4/10	133/107	23/53	84/63	112/12	53/131
P030	13741	14439	43/30	4/9	70/106	22/51	82/35	107/12	0/128
P031	14440	15249	43/30	4/9	71/107	44/51	83/62	112/12	52/130
P033	15250	15986	43/30	4/10	73/23	46/52	84/63	112/13	53/131

A.5. Post-norm Vs. Pre-norm Transformer Block

The transformer architecture, a foundation of modern natural language processing and computer vision tasks, exhibits two primary variants in its block design: post-norm and pre-norm. These variants differ in the placement of layer normalization within the transformer block structure.

In the post-norm configuration, layer normalization is applied after the multi-head attention and feedforward operations. Conversely, the pre-norm design positions the normalization layer before these operations. Figure A.1 illustrates these two configurations.

While both variants are widely used in various implementations, recent research has been conducted to evaluate the advantages of one over the other. Several empirical [32, 75] and theoretical [92] studies advocate for a pre-norm design in Transformer architectures. This earlier placement of normalization layers has been shown to stabilize the training process and potentially enhance performance. However, it is important to note that despite these findings, both post-norm and pre-norm designs continue to be employed in practice.

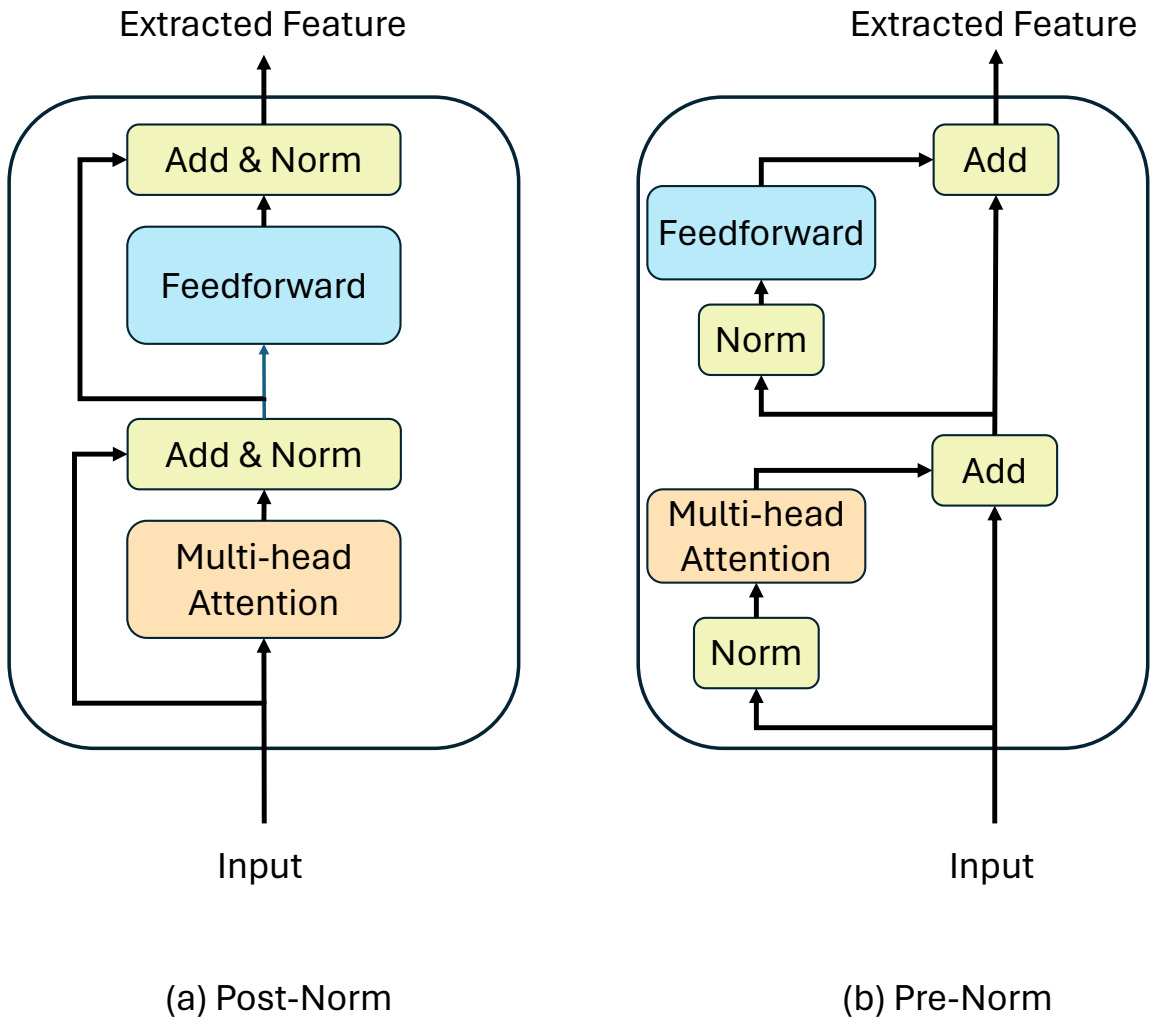


Figure A.1: Transformer encoder block variants. (a) Post-norm: normalization applied after multi-head attention and feedforward operations. (b) Pre-norm: normalization applied before multi-head attention and feedforward operations.

A.6. Visualization of Attention

The process begins by loading the trained model and registering forward hooks to capture the output of specific layers. In PyTorch, hooks are a mechanism that allows for the interception and modification of tensors during the forward or backward pass of the neural network. In this case, forward hooks are utilized to extract intermediate representations. A hook on the QKV (Query-Key-Value) computation layer to extract the combined QKV tensor is registered.

After registering the hook, the model is fed with the input data. As the data flows through the network, the registered hook capture the required intermediate representations.

The QKV tensor, obtained via the hook, is then split into its constituent parts: query (Q) and key (K). These are reshaped to incorporate the multi-head attention structure. The attention weights are computed by performing a matrix multiplication between Q and the transpose of K, resulting in a tensor that represents the attention scores for each head across all patches of the input image.

To create a single attention map, the attention weights are summed across all heads. This aggregated attention map is then reshaped and upsampled to match the dimensions of the original size of the input frames of the periocular recordings. The resulting heat-map is overlaid on the first frame of the recording. The following pseudo-code illustrates the key steps in this process, including the hook registration:

```
def register_hooks(model):
    def qkv_hook(module, input, output):
        hooks_output['qkv'] = output.detach()
    model.qkv_layer.register_forward_hook(qkv_hook)

def generate_attention_heatmap(model, input_data):
    hooks_output = {}
    register_hooks(model)

    _ = model(input_data)  # Forward pass, hooks capture outputs

    qkv = hooks_output['qkv']
    q, k = split_and_reshape_qkv(qkv)

    attention_weights = compute_attention(q, k)
    attention_map = aggregate_attention(attention_weights)

    upsampled_map = upsample(attention_map, target_size)

    overlay_heatmap(upsampled_map, input_data[0])
    display_heatmap()
```