Emotion Recognition in Adaptive Virtual Reality Settings

Emotion Recognition in Adaptive Virtual Reality Settings: Challenges and Opportunities

Seyed Muhammad Hossein Mousavi[1,*,†], Bulat Khaertdinov[2,†], Pedro Jeuris[2], Enrique Hortal[2], Davide Andreoletti[1] and Silvia Giordano[1]

[1]University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland
[2]Maastricht University,Maastricht, Netherlands

Abstract

Recently, there has been a notable surge of interest in Emotion Recognition (ER) systems, primarily due to their potential in improving interactions between humans and computers. Meanwhile, Virtual Reality (VR) has emerged as a groundbreaking technology that is also capable of transforming human-computer interaction through the simulation of immersive and flexible environments. The integration of ER into VR holds great promise for further advancing human-computer interaction by allowing the virtual environment to adapt to the user's emotional state. This adaptive VR setting is particularly relevant in fields such as education and gaming, where there is often the need to adapt the content to a person's emotions. However, applying traditional ER systems to adaptive VR settings comes with several challenges. In this paper, we identify the key differences between traditional ER and ER performed in VR environments. Specifically, we argue that the two scenarios primarily differ in terms of data collection methodologies and handling multimodality. After reviewing the main modalities considered in ER, and describing existing datasets, we delve into the challenges associated with these factors, highlighting the limitations of using traditional datasets in adaptive VR settings, and the fact that traditional ER models are not designed to effectively handle the multiple modalities arising from the VR setting. In addition to discussing these challenges, we also explore unique opportunities that arise from overcoming them. These opportunities include acquiring diverse datasets, eliciting genuine emotional responses, and exploiting multiple data modalities.

Keywords

Virtual reality, emotion recognition, datasets, speech, body tracking, bio-measurements

1. Introduction

Emotions are complex physiological and psychological states provoked by a variety of situations and stimuli. They play an essential role in human communication and are expressed through various modalities, such as facial and vocal expressions, body language, and diverse physiological responses. Emotion Recognition (ER) is the task of detecting emotions from a set of observed

Workshop on Advances of Mobile and Wearable Biometrics, MobileHCI'23, September 26, 2023, Athens, Greece
*Corresponding author.
†These authors contributed equally.
seyed.mousavi@supsi.ch (S. M. H. Mousavi); b.khaertdinov@maastrichtuniversity.nl (B. Khaertdinov); p.jeuris@student.maastrichtuniversity.nl (P. Jeuris); enrique.hortal@maastrichtuniversity.nl (E. Hortal); davide.andreoletti@supsi.ch (D. Andreoletti); silvia.giordano@supsi.ch (S. Giordano)
0000-0001-6906-2152 (S. M. H. Mousavi); 0000-0003-1651-0657 (B. Khaertdinov); 0009-0004-2471-4654 (P. Jeuris); 0000-0003-2119-4169 (E. Hortal); 0000-0003-3790-9341 (D. Andreoletti); 0000-0003-2603-9029 (S. Giordano)

modalities. ER has drawn significant attention in recent years thanks to the latest advances in sensing technologies and Artificial Intelligence. In particular, ER models based on Machine Learning and Deep Learning have shown tremendous success, especially when using multimodal data containing facial expressions and audio.

ER has the potential to enhance human-computer interaction, particularly in domains where emotions provide valuable insights into user experiences, such as education and gaming. Another notable technology that is gaining momentum in these fields is Virtual Reality (VR). This technology simulates virtual environments in which users experience unparalleled levels of immersion, and can actively interact with the virtual objects surrounding them. Moreover, in VR it is possible to flexibly modify the virtual environment, for example by dynamically changing the properties of the virtual objects. The integration of ER in VR settings makes it possible to implement immersive yet personalized experiences that adapt to the users based on their emotions, therefore further increasing their engagement. As an example, consider an individual learning how to use complex equipment by interacting with its VR digital twin. Based on the emotion detected by the ER system, the difficulty level can be adjusted accordingly. For instance, it can be reduced in case the person feels anxious or frustrated about performing certain actions, or it can be increased in case they feel bored.

Despite the recent advancements in ER systems, their applications in an adaptive VR setting is challenging, mostly due to i) the characteristics of the datasets traditionally collected and used to develop ER systems, and ii) the techniques implemented to handle multimodality. Specifically, emotions in existing datasets are generally not spontaneously evoked or acted (in some cases in a non-natural or exaggerated way), and they are usually annotated with labels (such as joy and fear) that are not relevant to common adaptive settings (such as adaptive learning). Furthermore, existing ER datasets are often collected from a limited number of individuals and do not consider a wide range of factors that influence the expression of emotions, such as cultural backgrounds, personality traits, and the specific activities in which individuals are engaged. This lack of diversity limits the generalizability of ER models across different individuals and VR scenarios, in which personalization and context-awareness are crucial. Then, data collected in VR are inherently multimodal, with certain modalities (or combinations thereof) being more relevant for specific use cases. Thus, unlike most state-of-the-art algorithms that are inclined towards some specific combination of modalities (e.g., facial expressions and audio signals), the ER models in VR should extract meaningful representations from any combination of available data and fuse them in a flexible manner.

However, VR also brings unique opportunities to go beyond the existing ER systems in terms of data collection and exploitation of multimodality. First, VR offers the possibility of reconstructing diverse scenarios that provoke genuine emotional responses and natural behavior during interaction. Besides, collecting synchronized data coming from multiple input modalities in such settings provide an opportunity for self-supervised and unsupervised frameworks to learn representations shared between modalities, including less commonly used data inputs, without using or with limited use of data annotations. In this paper, we thoroughly review the current landscape lying in the intersections of Emotion Recognition and adaptive Virtual Reality solutions. In particular, the contributions and outline of our paper are summarized as follows:

- The modalities used for emotion recognition and their applicability in VR settings along with the datasets commonly used in the ER research are introduced in Section 2.
- In Section 3, we introduce the challenges associated with re-using the current advances and datasets for ER. In particular, we aim to illustrate the identified challenges with the relevant examples from the literature and, where applicable, with the relevant experiments on open-source datasets.
- We formulate the opportunities for ER in adaptive VR in Section 4. In particular, we propose a novel way to collect multimodal data in the VR environments using annotations based on the theory of flow [1] and summarize the ideas that could extend the data collection protocols to be more user-centric and take into account user behavior. Besides, we present the opportunity of employing shared multimodal representation learning methods based on the current advances from other domains and highlight the aspects that have to be covered for a smooth adaptation of this paradigm.

2. Background on Emotion Recognition

2.1. Modalities

2.1.1. Bio-measurements

In recent years, particular attention in neuroscience research has been drawn by implicit feedback human bodies produce while experiencing certain emotions. In particular, bodily reactions to certain emotions include but are not limited to changes in heart rate and breathing tempo, different levels of sweat gland activities, and tension in muscles [2].

Electrodermal activity (EDA) signals, also known as Galvanic Skin Response (GSR), are recorded by sensors that measure changes in skin conductance which can be indicative of emotional arousal levels. Heart Rate Variability is another source of information that can be used to reflect the changes in the affective state. In particular, Electrocardiogram (ECG) is a robust heart rate monitor that requires connecting multiple electrodes to subjects' chests. A less intrusive and more lightweight measurement of heart rate variability is Blood Volume Pulse (BVP). Skin temperature (SKT) is also used for affect recognition, although it could be influenced by several factors and typically is a weaker signal compared to EDA and BVP [2]. Wearable devices, including both commercial-grade and research-grade options such as smart watches, bracelets, and rings, have the capability to record EDA, BVP, and SKT data through integrated electrodes.

Electroencephalography (EEG) signals, unlike the other physiological sensors, can be used for evaluating two-dimensional affect recognition, i.e. both arousal and valence levels. However, EEG devices require complex installation and, ideally, laboratory settings to collect accurate data.

While the described sensors can be exploited to recognize certain affective states, not all of them can be conveniently employed in real-life VR settings due to their intrusiveness, costs, and integration complexity. While most of the commercial sensors currently available in the market do not provide access to raw data, the research-grade devices are more expensive and do not necessarily provide convenient means for real-time integration. Besides, more complex

sensors, such as ECG and EEG, require a stationary environment during recording for obtaining high-quality data [3].

2.1.2. Body Movements

Body movement is represented as a sequence of joint positions and orientations over time. To obtain body movements, motion capture technologies, like the Kinect and VR headsets, are commonly used. These technologies mainly differ in the number of joints they track (e.g., the Kinect performs full-body tracking, while VR headsets only capture a subset of joints, generally head, and hands). The raw body movements are processed to extract features in various domains, specifically the time, frequency, and time-frequency domains. In the time domain, features such as velocity, acceleration, trajectory, and angle of joints [4], [5] are generally computed. Features computed in the frequency domain are particularly useful in detecting abrupt or fast movements (e.g., associated with emotions like anger or surprise), which might not be evident in the time domain. Examples of such features are the amplitude and phase of the Discrete Fourier Transform (DFT) [6]. As for the time-frequency domain, statistical features are generally computed from the wavelet transformations, such as discrete wavelets [7] and Gabor wavelets [8].

While body movements are relatively easy to capture, it is challenging to disambiguate emotions based solely on body movements. Hence, this modality is often used in combination with other signals (e.g., vocal cues) or used to tackle a more limited task, such as the identification of the polarity of the emotion (negative vs positive).

2.1.3. Audio

Speech, as an important modality to convey emotions, contains both semantic information and paralinguistic cues. The latter includes features such as intonation, timing, volume, and pitch and, unlike the former, can be transferred through speech only [9]. To work with speech for emotion recognition a suitable representation of these properties is needed. The Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10] is a low-level descriptor often used for ER based on speech. This descriptor extracts 88 features related to acoustic information. However, in eGeMAPS, semantic information is lost. Another commonly used features are Mel-frequency Cepstral Coefficients (MFCCs). MFCCs have been widely, and successfully, used for automatic speech recognition as they contain information on the lower frequencies of speech and emulate how humans perceive sound. Thus, MFCCs contain both acoustic and semantic information about the speech. However, MFCCs are less robust to noise and ignore the spectrum phase [11], which includes temporal characteristics of phonetic transitions. A third option is to use self-supervised features extracted by neural network models such as wav2vec2.0 [12] or HuBERT [13]. Pre-trained wav2vec2.0 models have already shown better performance compared to eGeMAPS and MFCCs in the task of speech ER [14, 15] but come with a higher computational cost. Besides, a combination of wav2vec2.0 features with eGeMAPS can positively influence the classification performance for ER [14].

Compared to other modalities, such as facial expressions, speech is easy to acquire in a VR context where the headset may introduce visual occlusions [16]. However, not all VR scenarios

require a user to speak, which makes it a limitation of this modality.

2.1.4. Facial Expression

Facial expressions are highly correlated with a person's emotional state [17]. Facial expressions can be captured using various instruments, such as cameras, depth sensors, and wearable devices. Each of these instruments tracks facial expressions differently. Cameras capture the overall facial movements, including the eyes, eyebrows, and mouth [18]. Depth sensing technologies capture the three-dimensional structure of the user's face [19]. Wearable devices, such as ElectroMyoGraphy (EMG) sensors, measure the electrical signals generated by facial muscles.

Generally, several pre-processing operations are performed on the captured signals. These operations depend on the type of sensors used for data collection. For instance, in data collected with camera sensors, background removal and adjustment of image intensity value are common operations. Features for downstream machine learning models are generally computed from the spatial and frequency domains. Examples of features computed in the spatial domain are the Local Binary Patterns (LBP) [20], which are very robust against changes in illumination levels [21], and the Histogram of Oriented Gradients (HOG) [22], which are robust against geometric and photometric transformations. Examples of features computed in the frequency domain are the Local Phase Quantization (LPQ) [23] and the Gabor Filters [24], which are very robust against blurring effects and change in rotation, resizing, and illumination, respectively. While facial expressions allow for effective emotion recognition, their recognition in a VR setting is problematic. In fact, cumbersome infrastructure with particular equipment like specific camera sensors and lighting is required. Additionally, an important part of the face is obscured as its upper part is covered by the VR headset. On the other hand, the lower part can be detected by external cameras or even cameras embedded in the VR headset itself.

2.1.5. Pupil Dilation and Eye Gaze

Eye movement signals have proved correlated with a person's emotional state [25]. In particular, high pupil dilation is a sign of emotional arousal [26]. To a lower extent, also features related to gaze (e.g., fixation) provide useful information about emotional states, but are generally used as a complementary modality [27]. Pupil dilation and gaze information can be captured using eye-tracking technologies, which are nowadays often embedded into VR headsets. These technologies allow accurately capturing changes in pupil size [28], as well as information about viewing directions (such as pitch and yaw angles). Some common pre-processing techniques applied to pupil dilation signals include blink and saccades (i.e., rapid eye movements) elimination [29], and the use of wavelet transformations for noise removal [30]. Then, statistical features are generally derived from the time series of the pupil diameter [31]. The analysis of pupil dilation is highly effective in ER, especially to tackle the task of arousal estimation. However, this signal also comes with several limitations. In particular, pupil dilation is strongly affected by lighting conditions, underlying cognitive processes, and individual differences (e.g., age and health status). Moreover, it is challenging to disambiguate emotions with the same level of arousal, but with different levels of valence (such as excitement and anger) solely based on pupil dilation data [32].

2.2. Datasets

In this section, we provide a concise overview of the main datasets currently available for ER that encompass the mentioned modalities. Our categorization of datasets is based on two main criteria: whether they were gathered specifically for VR environments or not, and whether the data they encompass is unimodal or multimodal. The main characteristics of each dataset, such as employed modalities and elicitation media, are summarized in Table 1.

Dataset	Modalities	EM	Annotations	subjects	VR context
WESAD [33]	Wrist: BVP, EDA,TEMP, and ACC Chest: ECG, EDA, EMG, and TEMP	Baseline, stress, amusement	Stimuli, self (limited)	15	✗
PacoLab [34]	Body Motion Joints	Anger, Sadness, Happiness, and Neutral Emotions	Stimuli, self (limited)	30	✗
IKFDB [19]	Image Frames (Color and Depth)	7 Facial Expressions	Stimuli, self (limited)	40	✗
EMOVOCorpus [35]	Audio	6 Vocal Expressions	Stimuli, self (limited)	6	✗
AMIGOS [36]	Body Motion, Facial Expressions, Color and Depth EEG, ECG, GSR	7 Main Expressions	Video	40	✗
DEAP [37]	EEG, GSR, Respiration Amplitude, TEMP, BVP, EMG, EOG, Facial Expressions	Arousal and Valence	Music Video	22	✗
MAHNOB-HCI [38]	videos, audio, eye gaze, EEG, and peripheral/ physiological signals.	Arousal and Valence, dominance, and predictability	Self	27	✗
CASE [39]	ECG, BVP, EMG, EDA, TEMP, respiration sensors	Arousal and Valence	Self	30	✗
IEMOCAP [40]	Speech, Head & face Motion capture	Emotions	Stimuli	10	✗
K-EmoCon [41]	Wrist: BVP, EDA, TEMP, ACC, HR, IBI Other: Speech, EEG, ECG, Attention & Meditation	Arousal/Valance, Emotions, BROMP	Self, External, Partner	up to 32	✗
RAVDESS [42]	Speech, Video	Emotions	External	24	✗
CREMA-D [43]	Speech, Audio	Emotions	External	91	✗
Marín-Morales et al. [44]	EEG, ECG, Heart Rate Variability (HRV)	Arousal and Valence	360 VR Videos	60	✓
VREED [45]	ECG, GSE, Gaze Tracking	Arousal and Valence	360 VR Videos	34	✓
Dozio et al. [46]	Image	7 Facial Expressions	360 VR Videos	75	✓
DER-VREED [47]	EEG	Happiness, Fear Calmness, and Boredom	360 VR Videos	32	✓
CEAP-360VR [48]	Pupil dilation, head movements, eye gaze EDA, SKT, TEMP, Acc, HR, IBI	Arousal and Valence	Continuous self-annotation	32	✓

Table 1
Overview of the widely-used open-source emotion recognition datasets.

2.2.1. Unimodal Datasets for VR settings

A Dataset for Emotion Recognition using Virtual Reality and EEG (DER-VREEG) [47] contains physiological EEG signals annotated with 4 classes of emotions namely happiness, fear, calmness, and boredom. The dataset has been collected with the help of 32 participants, using Interaxon Muse 2016 to measure the brainwave signals from four different channels (namely, AF7, AF8, TP9, TP10). The Alpha, Beta, Delta, Theta, and Gamma bands are then extracted from the raw signals. Emotion elicitation has been performed by projecting videos in VR. The dataset presented in [44] contains EEG, ECG, and Heart Rate Variability (HRV) signals annotated with arousal and valence measures, for 60 individuals. Signals have been measured with a Samsung Gear VR HMD, which has also been used to elicit emotions by projecting 360-degree panoramic views.

The dataset presented in [46] contains facial expressions annotated with seven classes of emotions (namely joy, fear, disgust, surprise, sadness, anger, and neutral expressions), which have been collected from 75 participants. Emotions have been evoked by means of 360-degree audiovisual content, projected using various types of VR headsets.

2.2.2. Multimodal Datasets for VR settings

The Continuous Physiological and Behavioral Emotion Annotation Dataset for 360-degree videos (CEAP-360VR) [48] consists of head and eye movement, pupil dilation, and physiological signals (EDA, SKT, BVP, Acc, and heart rate). The dataset has been collected from 32 participants, and emotions have been evoked using short 360-degree videos projected in VR. Such videos have been selected from the open database of stimuli videos [49], and are annotated with arousal and valence levels. Nevertheless, the labels in this dataset have been provided by the participants themselves by means of self-annotations of arousal and valence scores. The physiological measurements have been collected through the Empatica E4 wristband, whereas the VR headset employed is HTC VIVE Pro Eye HMD. The VR Eyes Emotions Dataset (VREED) dataset [45] contains eye tracking and psychological signals (ECG and GSR) annotated in arousal and valence from 34 participants. The Biopac MP 150 system was used to continuously acquire ECG, GSR, and eye-tracking signals. Emotion elicitation has been conducted by immersive 360-degree videos using a VR headset.

2.2.3. Datasets for Emotion Recognition in non-VR settings

Examples of unimodal datasets are the EMOVO Corpus [35], IKFDB [19] and PACO Lab [34], which consist of speech signals, facial expressions, and body motions, respectively. The emotion elicitation means employed are text (for the EMOVO Corpus) and self-stimuli (for the other datasets). As far as multimodal datasets are concerned, the K-EmoCon [41] dataset consists of videos, speech, bio-measurements from wrist-worn (e.g., EDA and heart rate), head-worn (e.g., EEG) and chest-worn (e.g., ECG) sensors. Emotions are evoked by letting participants discuss a political topic. The DEAP dataset [37] contains facial expressions and physiological signals (respiration amplitude, blood volume, and others). Videos and music have been used for emotion elicitation. The AMIGOS dataset [36] contains body motion, facial expressions, and physiological signals (e.g., EEG, and ECG). Emotion elicitation has been performed using videos. The IEMOCAP dataset [40] contains speech and motions, and emotions have been evoked by talking to the participants. The WESAD dataset [33] contains physiological signals (e.g., EDA) and accelerations. In this dataset, a combination of emotion elicitation strategies have been implemented, such as stress tests, videos, and guided exercises.

3. Challenging Aspects of Emotion Recognition in Adaptive VR Systems

In this section, we elaborate on the main challenges that prevent the application of existing ER models to the adaptive VR setting. Specifically, we discuss challenges related to data collection and the handling of multimodality.

3.1. Data Collection

3.1.1. Emotion Elicitation

Effective emotion elicitation is essential in constructing reliable datasets for training ER algorithms. Specifically, the observed signals (e.g., body motion and voice cues) should align with the input stimuli. However, controlling the emotion elicitation process is challenging due to several factors. Firstly, there is a lack of comprehensive understanding regarding the key factors that trigger emotions. Additionally, individual subjectivity plays a role, as each person may react differently to the same emotional stimulus. Furthermore, an individual's psychological state can influence the perceived emotion, regardless of the input emotional stimulus. This lack of control over emotion elicitation poses a significant challenge to the collection of reliable datasets. In fact, as further explained in subsections 3.1.2 and 3.1.4, failing to induce the target emotion may impede the correct annotation of the collected data, which in turn compromises its validity for developing ER systems.

To illustrate the mentioned challenge of emotion elicitation we draw an example from the widely used WESAD dataset [33] that provides a three-class classification problem (baseline-stress-amusement).

Specifically, we implemented a one-dimensional 3-layer CNN widely used as a backbone model in related works [50, 51, 52]. As input, all biomarker modalities (EDA, BVP, and TEMP) are resampled to the same frequency, segmented into 10-second intervals, and stacked to create a single input with three channels. The flattened embeddings of the CNN layers are then passed through a linear classification head. The model is trained using the whole dataset, i.e. all subjects, to obtain the CNN representations. In Figure 1, the t-SNE projections of the embeddings are visualized. As can be seen, multiple dense clusters of amusement instances appear in the middle of the baseline class embeddings. In other words, even when all the available subjects are used for training, the model fails to separate instances of amusement and baseline classes in most of the cases showing that the used stimuli (i.e. humorous videos) have not been successful to provoke higher levels of arousal typical for the amusement state. On the other hand, the stress class is quite well separable from the baseline emotion. The stressful state, in this dataset, has been evoked by the Trier Social Stress Test [53].

3.1.2. Data Annotation

Data annotation can be approached in various ways, each with its own set of challenges. One approach involves associating the user's emotion with the emotional stimulus used to trigger it. For example, if a person is exposed to a stimulus intended to induce happiness, the corresponding data is labeled as belonging to the happiness class. However, as illustrated in Section 3.1.1, there may be instances where the user's emotional response does not align with the expected emotion, resulting in a distorted dataset. A second approach is self-annotated datasets. This approach consists in asking users to express their own emotions. However, users may lack awareness of their emotions or provide incorrect labels, leading to biased datasets. A third option for data annotation involves a group of observers estimating the user's emotions. However, even experienced individuals may find the task of accurately recognizing and assessing emotions challenging, introducing another potential source of bias in the dataset [41].

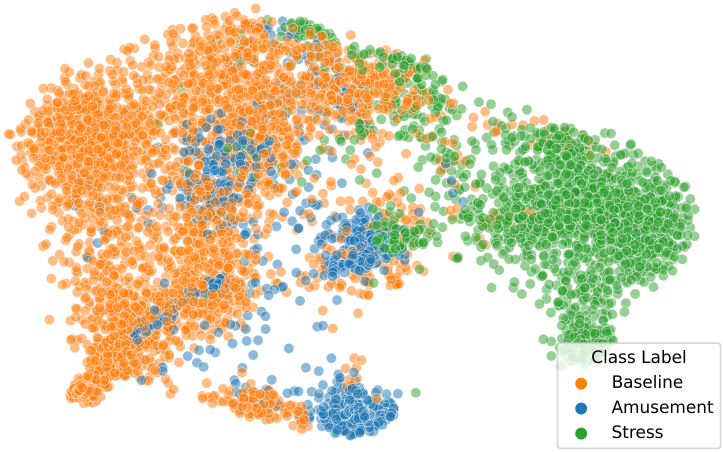

Figure 1: WESAD embeddings projected on the two-dimensional space using t-SNE. The embeddings were obtained using 3-layer 1D-CNN processing 10 seconds intervals.

3.1.3. Emotional Model

Emotion Recognition systems in adaptive scenarios, besides classical emotional models, require more specific affective models measuring user engagement. In particular, it is important to detect the frustration levels of users to either collect feedback about certain scenarios or adapt the experience to increase user engagement [54].

One of the challenges of re-using open-source datasets to train models for adaptive VR settings is that most of these datasets only contain more common emotional models, such as the discrete Ekmanian model with six basic emotions [55] or dimensional models describing levels of arousal and valence [56]. From the datasets covered in this paper, only one dataset, namely K-EmoCon [41], contains, among others, an emotional model that can be used to track the engagement level of users. Specifically, the dataset authors also provide the so-called BROMP annotations [57] that are specifically used in education. Nevertheless, the class representation for this type of annotation is extremely imbalanced which makes it difficult to be used as a basis for training a classification model.

It is important to mention that the relevant literature also suggests a mapping between common emotional models and engagement-related ones. For example, in [58] and [59], authors provide a mapping between a two-dimensional model and discrete emotions related to the theory of flow based on empirical data they collected using educational and computer gaming scenarios, respectively. Nevertheless, most of the datasets have been collected with the goal to elicit certain emotions and they are not tailored to various engagement signals. Besides, the affective states and their ranges vary significantly for different domains depending on certain education or gaming scenarios, meaning that data collection covering various scenarios is preferred.

3.1.4. User and Task Heterogeneity

The expression of emotions exhibits a certain level of consistency among individuals, but each person also has a unique way of expressing emotions. This individuality poses a challenge in developing a universal system that can effectively analyze the emotions of diverse individuals. The difficulty primarily stems from the subjective nature of emotions and the influence of cultural factors. Regarding the former, individuals may respond differently to the same stimulus. For instance, some individuals may not exhibit outward signs of a specific emotion, while others may express it in unconventional ways. As for the latter, cultural norms and expectations shape how emotions are expressed and interpreted. Facial expressions or vocal cues that convey a particular emotion in one culture may carry different meanings or interpretations in another culture.

While the problem of user heterogeneity is widely recognized in the field of affective computing [60], the issue of task heterogeneity in emotion recognition also deserves attention. Indeed, how individuals express emotions is influenced by the type of task they are engaged in. In the context of VR, this problem becomes particularly significant due to the versatility offered by this technology, which enables users to immerse themselves in various types of tasks, such as operating industrial machinery or engaging in educational assignments. Consequently, it is challenging to extend emotion recognition algorithms, previously trained on datasets collected from individuals performing specific tasks, to effectively work on individuals performing different tasks.

In order to visualize the problem of user heterogeneity, we employ a Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol on unimodal and multimodal datasets, WESAD [33] and K-EmoCon [41], respectively. The LOSO-CV protocol consecutively uses data from each subject in a dataset as a test set. The remaining subjects are exploited for training and validation. For both datasets, we train a 3-layer CNN that we have previously introduced in Section 3.1.1 for the biomarker data. Additionally, for the audio modality in K-EmoCon, a single dense layer was trained on top of the eGeMAPS features. We visualize the distribution of model performance across different subjects in LOSO-CV in Figure 2. As can be seen, the performance spread is wide for both datasets. Besides the fact that the emotions were perceived differently, such performance could have also been caused by the elicitation protocols that did not trigger similar affective states in different subjects.

3.2. Multimodality

3.2.1. Unsupervised Representation Learning

The last couple of years have seen rapid developments in representation learning in the major AI areas, including Computer Vision, Natural Language Processing (NLP), and Automatic Speech Recognition. Novel Self-Supervised Learning (SSL) frameworks have been extensively used to pre-train large Deep Learning models to extract robust and meaningful features from different types of data. Typically, the state-of-the-art SSL approaches leverage large unlabeled datasets to pre-train the models on so-called pre-text tasks. These tasks formulate auxiliary objectives, such as predicting part of the input signal from the surrounding context [61] or learning distortion-invariant features [62], that allow the models to extract meaningful features

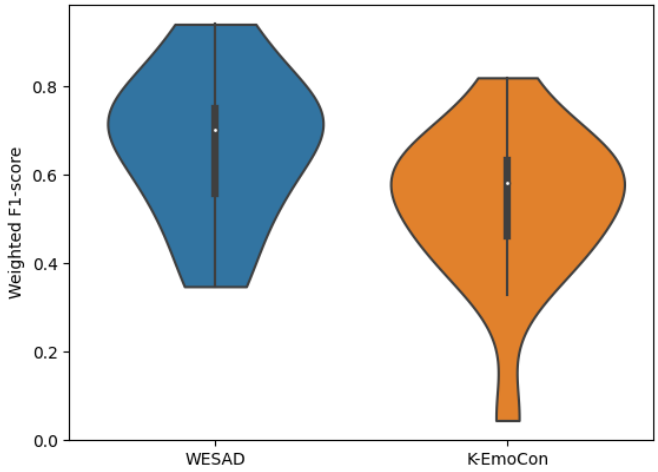

Figure 2: Distribution of weighted F1-scores for LOSO cross-validation for the WESAD and K-EmoCon datasets.

from data.

SSL paradigm has been also applied to emotion recognition to learn unimodal features from different modalities. Experiments conducted on various modalities, such as ECG data [63], facial videos, text, and audio data [64], and EDA, BVP, and SKT data [50], show that unsupervised representation learning is a promising direction that allows Deep Learning models to learn robust unimodal representations without acquiring data annotations. Another approach that has been used for audio and video modalities is to exploit the large speech models, such as wav2vec2.0 [12, 14] and HuBERT [13] for audio, pre-trained on huge unimodal datasets. A similar approach has been used for physiological data in [65] where researchers exploited a large private unannotated dataset to pre-train the Deep Learning encoder.

Given the significant progress in exploiting SSL for unimodal emotion recognition, there are still some challenging aspects when adapting them to VR settings. First, the publicly available unimodal datasets are quite limited for certain modalities relevant to the VR context, such as physiological data and pupil dilation. Besides, pre-trained models do not guarantee optimal performance when transferring representations to emotion recognition tasks. To demonstrate this problem we compare a large pre-trained speech model, namely wav2vec2.0 [12], adapted to emotion recognition as proposed in [14] and a simple linear encoder built on top of the handcrafted eGeMAPS features on two datasets, IEMOCAP [40] and K-EmoCon [41]. As can be seen from the F1-score values presented in Table 2, the pre-trained wav2vec2.0 features perform significantly better compared to the encoder for eGeMAPS which was trained from scratch on IEMOCAP. However, for K-EmoCon, eGeMAPS features passed through a simple linear model are more robust compared to the features extracted by wav2vec2.0.

3.2.2. Flexible Multimodal Fusion

As discussed in subsection 2.1, emotion recognition can be approached by processing data from multiple modalities. Moreover, in the last decade, researchers have paid a significant

Features	IEMOCAP	K-EmoCon
wav2vec2.0	63.2%	31.4%
eGeMaps	48.6%	37.0%

Table 2
Average F1-score for IEMOCAP and K-EmoCon.

amount of attention to multimodal emotion recognition using a variety of input modalities simultaneously. While each of these data sources can contain informative and unique cues about the experienced affective state, each of them also brings its specific limitations. For instance, algorithms based on facial expression data are vulnerable to different types of occlusions, such as the VR headset. What is more, in certain settings and tasks, some data inputs can be irrelevant. In particular, audio data might be the main source of information when the scenario implies oral communication, whereas, in some gaming and educational scenarios, subjects may not be supposed to communicate orally. In such cases, flexible fusion algorithms are needed to choose the most relevant cues to evaluate the affective states of the subjects.

Another challenging aspect that has to be considered is the predictive power of certain modalities. For example, the literature highlights that most of the sensors that can be installed in simple non-intrusive wearable devices, such as EDA, BVP, and SKT sensors, collect signals that could explain various intensity levels of emotions, or arousal levels. However, there is no evidence that changes in valence levels, or pleasantness of emotions, cannot be tracked using the mentioned modalities [66]. Hence, the choice of modalities to be used should also be tailored for a specific scenario and downstream task.

Finally, selecting the type of fusion algorithm is a challenging process. While the late decision-level fusion techniques are more lightweight and do not require data streams perfectly aligned between modalities, they do not take into account the inter-modal correlations [67]. On the contrary, the early and feature-level fusion methods employ the inter-modal dependencies in the representation learning and prediction-making process. Nevertheless, feature fusion, which is typically done through feature concatenation, requires well-synchronized inputs and is less flexible in terms of the set of input modalities.

4. Opportunities for Emotion Recognition in VR

4.1. Data Collection

4.1.1. Collecting Datasets for Emotion Recognition in Adaptive VR

VR holds significant potential for collecting extensive datasets for ER in adaptive scenarios. This potential primarily arises from the ability to track users' signals across different modalities (e.g., body motion and physiological signals) and evoke stronger and more genuine emotional reactions. Furthermore, VR allows for precise control over environmental factors like colors and shapes, further enhancing the emotional experience. However, as discussed in subsection 3.1.1, the lack of complete understanding of which factors actually trigger emotions limits the control over emotion elicitation. To address this challenge, authors in [68] propose leveraging the theory of flow [1] as a theoretical guideline for systematic emotion elicitation.

Given that the theory of flow establishes a connection between task difficulty, individual skills, and emotional response, we argue that it is particularly well-suited for eliciting emotions in an adaptive VR setting, for the following reasons. Firstly, in most adaptive VR settings, such as those found in education and training scenarios, users are often required to perform tasks that can be adjusted in difficulty based on the users' skills. Following the theory of flow, when the difficulty level is customized to match the skills of the users, it can induce the following emotions: anxiety if the difficulty exceeds the users' skills, engagement if the difficulty aligns well with the users' skills, and boredom if the difficulty is below the users' skills level. Secondly, as discussed in subsection 3.1.3, traditional emotional models tend to focus on emotions (e.g., joy and fear), which may not properly align with most adaptive VR scenarios. In these scenarios, the emotions identified by the theory of flow, namely anxiety, engagement, and boredom, are more relevant and meaningful in understanding the user's emotional experience.

Based on these considerations, the authors of Ref. [68] presented a VR application called the Magic Xroom, which utilizes the theory of flow for precise control over emotion elicitation. The main objective of the Magic Xroom is to collect extensive datasets for ER in adaptive VR scenarios. In the Magic Xroom, the task difficulty dynamically adjusts based on participants' skill levels, aiming to strike an optimal balance that induces the desired emotional states of engagement, anxiety, or boredom. Continuous monitoring of participants' body responses, including body movements, heart rate, and skin conductance, is performed to gather multimodal data. Additionally, users can express their emotions through a virtual panel, providing explicit feedback on their emotional experiences. By adopting a systemic approach to emotion elicitation and collecting multimodal data with relative annotations, the Magic Xroom offers a scalable system for collecting datasets to develop emotion recognition systems, particularly well-suited for the adaptive VR setting.

4.1.2. Personalized Data Collection

As discussed in subsection 3.1.4, developing universal ER systems that effectively work for different individuals is extremely challenging. However, it is possible to address this challenge by taking a different approach. In particular, instead of focusing on building universal systems, the emphasis can be placed on creating personalized datasets and constructing custom models. VR opens up possibilities for achieving this goal, for the following reasons. Firstly, VR is becoming increasingly used on a regular basis. For instance, individuals can engage in educational or entertainment scenarios that span multiple sessions. This provides an opportunity to observe individuals across these sessions and gather more data about them. Secondly, VR enables the flexible creation of various contexts, allowing one to observe individuals engaged in different tasks and situations. By having a significant amount of data from diverse contexts, it becomes possible to study emotional expressions more extensively, and better capture an individual's unique style of expression. Thirdly, leveraging the interactive nature of VR, users can provide instant emotional feedback through virtual panels or other means. This facilitates the annotation of data tailored to each specific individual, which further enhances the customization of datasets for building personalized ER models.

4.1.3. Behavior-Driven Emotion Recognition

Emotions are not solely discernible through body signals, such as voice and physiological data, but can also be inferred from behavioral clues (e.g., how an individual interacts with the objects surrounding her). In fact, the utilization of user behavior holds significant potential in revolutionizing emotion recognition algorithms. By providing a realistic, immersive, and adaptable environment where users can interact with virtual scenes, VR offers a remarkable opportunity to study the relationship between behavior and emotions that, to our knowledge, has never been tackled in previous works. Indeed, by integrating behavioral data with traditional signals, more comprehensive and accurate models can be developed. This novel approach opens avenues for capturing nuanced emotional experiences and enriching the understanding of the diverse ways in which emotions manifest.

4.2. Shared Multimodal Representation Learning

In recent years, Self-Supervised Learning or unsupervised representation learning has drawn a lot of attention in the research literature. Specifically, this paradigm offers an opportunity to pre-train large Deep Learning models using vast amounts of unannotated data, and later fine-tune the models to a certain downstream task using significantly smaller datasets. As previously described in Section 3.2.1, SSL has already been adapted to emotion recognition in two forms, namely unimodal SSL pre-training and the use of large models pre-trained on other tasks exploiting the same modalities.

The latest works in other major AI domains demonstrate that multimodal pre-training and representation alignment across modalities show superior results for both unimodal and multimodal downstream tasks [69, 70, 71, 72]. To the best of our knowledge, this direction has not been explored in the research regarding using various combinations of modalities related to the VR settings and for emotion recognition, in general, except for audio-visual cues [73]. Multimodal representation learning brings opportunities to improve the representations of affective states and align them between various modalities. This is especially promising for the adaptive settings, as SSL models typically have a better ability to transfer knowledge between various tasks. Moreover, unsupervised representation learning methods utilize discriminatory information, patterns, and correlations identified in multimodal data, whereas supervised methods are guided by strong priors from annotations. Given the challenge of annotating input data, the labels used by supervised models can be misleading and limited to a certain affective model. Besides, this approach has the potential to address the challenge of flexible multimodal fusion given that latent representations from different modalities are aligned.

Multimodal unsupervised representation learning typically requires large amounts of data. Hence, this would require the collection of large unannotated datasets with relevant synchronized modalities in an immersive environment covering various types of tasks eliciting different types of affective states as described in Section 4.1.1. Besides, another aspect to consider is that multimodal alignment might have to follow a certain schema to extract specific features related to different emotion dimensions given the limitations of some modalities in recognizing changes across certain dimensions (e.g., EDA signals and valence levels). The study exploring this direction has been recently conducted for an audio modality to disentangle the latent

representations for arousal, valence, and dominance dimensions [74].

5. Conclusion

Adaptive Virtual Reality, with its immersive and interactive nature, presents an opportunity to personalize the user experience based on their emotions with the goal to increase engagement and user satisfaction. These properties of the interactive sessions are crucial in learning and gaming environments to maximize the progress of users. However, there are notable differences between the settings required to build classical Emotion Recognition models compared to the ER algorithms suitable for adaptive VR environments.

In this paper, we aim to critically review the current landscape of ER and its suitability for adaptive VR settings. In particular, we introduce the modalities for ER and highlight the limitations that can arise when integrating them into the VR settings. We also summarize and describe the open-source datasets frequently used for unimodal and multimodal ER and focus on emotions elicitation and annotation protocols. Based on our observations and related literature, we formulate a list of challenges for adapting ER to adaptive VR and categorize them into two groups. First, data collection-related challenges contain crucial yet rigorous aspects of successfully eliciting genuine and spontaneous emotions and correctly annotating them. Besides, adaptive VR settings can benefit from certain emotional models that, to the best of our knowledge, are not employed in open-source datasets. Another category of challenges is associated with tackling multimodality. The current state-of-the-art ER models are typically based on two major modalities, namely facial expressions and audio, that cannot be dominant in VR applications. Thus, given that the largest datasets and pre-trained models are tailored for this combination of modalities, one of the challenges is to obtain meaningful representations for less common modalities that can be integrated into VR environments. Besides, another challenging aspect is to allow the flexible fusion of various modalities based on the selected VR scenario.

Finally, our paper suggests a set of opportunities that can be seen as a call for action in future studies on the intersection of ER and adaptive VR applications. First, we propose to cover the challenges associated with data collection. In particular, we present a novel setup for collecting data within the adaptive user-centric VR environment. Besides, we propose to perform personalized data collection and extract cues related to the behavior of individuals in order to adapt the models to various users smoothly. Finally, we propose ideas to adapt multimodal representation learning to ER in order to create robust shared representations and make multimodal fusion more flexible.

References

[1] M. Csikszentmihalyi, Beyond boredom and anxiety., Jossey-bass, 2000.

[2] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: First steps towards an automatic system, in: Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings, Springer, 2004, pp. 36–48.

[3] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, Electronic Notes in Theoretical Computer Science 343 (2019) 35–55.
[4] M. M. Del Viva, M. C. Morrone, Motion analysis by feature tracking, Vision research 38 (1998) 3633–3653.
[5] W. Liang, F. Wang, A. Fan, W. Zhao, W. Yao, P. Yang, Extended application of inertial measurement units in biomechanics: From activity recognition to force estimation, Sensors 23 (2023) 4229.
[6] B. Li, C. Zhu, S. Li, T. Zhu, Identifying emotions from non-contact gaits information based on microsoft kinects, IEEE Transactions on Affective Computing 9 (2016) 585–591.
[7] S.-L. Chang, C.-C. Hsu, T.-C. Lu, T.-H. Wang, Human body tracking based on discrete wavelet transform, in: Proceedings of the 2007 WSEAS International Conference on Circuits, Systems, Signal and Telecommunications, Citeseer, 2007, pp. 113–122.
[8] D. Vishwakarma, P. Rawat, R. Kapoor, Human activity recognition using gabor wavelet transform and ridgelet transform, Procedia Computer Science 57 (2015) 630–636.
[9] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2012) 99–117. doi:https://doi.org/10.1007/s10772-011-9125-1.
[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.
[11] M. Labied, A. Belangour, Automatic speech recognition features extraction techniques: A multi-criteria comparison, International Journal of Advanced Computer Science and Applications 12 (2021).
[12] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.
[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.
[14] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings, Proc. Interspeech 2021 (2021) 3400–3404.
[15] M. Macary, M. Tahon, Y. Estève, A. Rousseau, On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 373–380.
[16] M. Shah Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digital Signal Processing 110 (2021) 102951. URL: https://www.sciencedirect.com/science/article/pii/S1051200420302967. doi:https://doi.org/10.1016/j.dsp.2020.102951.
[17] P. Ekman, W. V. Friesen, Facial action coding system, Environmental Psychology & Nonverbal Behavior (1978).
[18] C. Vilchis, C. Perez-Guerrero, M. Mendez-Ruiz, M. Gonzalez-Mendoza, A survey on the pipeline evolution of facial capture and tracking for digital humans, Multimedia Systems (2023) 1–24.

[19] S. M. H. Mousavi, S. Y. Mirinezhad, Iranian kinect face database (ikfdb): a color-depth based face database collected by kinect v. 2 sensor, SN Applied Sciences 3 (2021) 19.
[20] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on kullback discrimination of distributions, in: Proceedings of 12th international conference on pattern recognition, volume 1, IEEE, 1994, pp. 582–585.
[21] B. Niu, Z. Gao, B. Guo, Facial expression recognition with lbp and orb features, Computational Intelligence and Neuroscience 2021 (2021) 1–10.
[22] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1, Ieee, 2005, pp. 886–893.
[23] J. Heikkila, V. Ojansivu, Methods for local phase quantization in blur-insensitive image analysis, in: 2009 International Workshop on Local and Non-Local Approximation in Image Processing, IEEE, 2009, pp. 104–111.
[24] V. Tadic, Z. Kiraly, P. Odry, Z. Trpovski, T. Loncar-Turukalo, Comparison of gabor filter bank and fuzzified gabor filter for license plate detection, Acta Polytechnica Hungarica 17 (2020) 1–21.
[25] J. Z. Lim, J. Mountstephens, J. Teo, Emotion recognition using eye-tracking: taxonomy, review and current challenges, Sensors 20 (2020) 2384.
[26] T. Partala, V. Surakka, Pupil size variation as an indication of affective processing, International journal of human-computer studies 59 (2003) 185–198.
[27] S. Wu, Z. Du, W. Li, D. Huang, Y. Wang, Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze, in: 2019 International Conference on Multimodal Interaction, ICMI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 40–48. URL: https://doi.org/10.1145/3340555.3353739. doi:10.1145/3340555.3353739.
[28] D. Martinez-Marquez, S. Pingali, K. Panuwatwanich, R. A. Stewart, S. Mohamed, Application of eye tracking technology in aviation, maritime, and construction industries: a systematic review, Sensors 21 (2021) 4289.
[29] M. E. Kret, E. E. Sjak-Shie, Preprocessing pupil size data: Guidelines and code, Behavior research methods 51 (2019) 1336–1342.
[30] C. K. Chui, G. Chen, et al., Kalman filtering, Springer, 2017.
[31] P. Ren, A. Barreto, Y. Gao, M. Adjouadi, Affective assessment by digital processing of the pupil diameter, IEEE Transactions on Affective computing 4 (2012) 2–14.
[32] M. Oliva, A. Anikin, Pupil dilation reflects the time course of emotion recognition in human vocalizations, Scientific reports 8 (2018) 4871.
[33] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing wesad, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM international conference on multimodal interaction, 2018, pp. 400–408.
[34] Y. Ma, H. M. Paterson, F. E. Pollick, A motion capture library for the study of identity, gender, and emotion perception from biological motion, Behavior research methods 38 (2006) 134–141.
[35] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, et al., Emovo corpus: an italian emotional speech database, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[36] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, I. Patras, Amigos: A dataset for affect, personality and mood research on individuals and groups, IEEE Transactions on Affective Computing 12 (2018) 479–493.

[37] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE transactions on affective computing 3 (2011) 18–31.

[38] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE transactions on affective computing 3 (2011) 42–55.

[39] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, F. Schwenker, A dataset of continuous affect annotations and physiological signals for emotion analysis, Scientific data 6 (2019) 196.

[40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359. doi:https://doi.org/10.1007/s10579-008-9076-6.

[41] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, U. Lee, K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, Scientific Data 7 (2020) 293.

[42] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018) e0196391.

[43] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (2014) 377–390.

[44] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, G. Valenza, Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors, Scientific reports 8 (2018) 13657.

[45] L. Tabbaa, R. Searle, S. M. Bafti, M. M. Hossain, J. Intarasisrisawat, M. Glancy, C. S. Ang, Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures, Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 5 (2021) 1–20.

[46] N. Dozio, F. Marcolin, G. W. Scurati, F. Nonis, L. Ulrich, E. Vezzetti, F. Ferrise, Development of an affective database made of interactive virtual environments, Scientific Reports 11 (2021) 1–10.

[47] N. S. Suhaimi, J. Mountstephens, J. Teo, A dataset for emotion recognition using virtual reality and eeg (der-vreeg): emotional state classification using low-cost wearable vr-eeg headsets, Big Data and Cognitive Computing 6 (2022) 16.

[48] T. Xue, A. El Ali, T. Zhang, G. Ding, P. Cesar, Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360 vr videos, IEEE Transactions on Multimedia (2021).

[49] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, L. M. Williams, A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures, Frontiers in Psychology 8 (2017). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02116. doi:10.3389/fpsyg.

2017.02116.
[50] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, S. Nanayakkara, Sigrep: Toward robust wearable emotion recognition with contrastive representation learning, IEEE Access 10 (2022) 18105–18120. doi:10.1109/ACCESS.2022.3149509.
[51] K. Matton, R. A. Lewis, J. Guttag, R. Picard, Contrastive learning of electrodermal activity representations for stress detection, in: NeurIPS 2022 Workshop on Learning from Time Series for Health, 2022. URL: https://openreview.net/forum?id=bSC_xo8VQ1b.
[52] G. Alhussein, M. Alkhodari, A. Khandokher, L. J. Hadjileontiadis, Emotional climate recognition in interactive conversational speech using deep learning, in: 2022 IEEE International Conference on Digital Health (ICDH), IEEE, 2022, pp. 96–103.
[53] C. Kirschbaum, K.-M. Pirke, D. H. Hellhammer, The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting, Neuropsychobiology 28 (1993) 76–81.
[54] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, B. Schuller, Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 2019, pp. 517–523. doi:10.1109/ACII.2019.8925464.
[55] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (1992) 169–200.
[56] L. A. Feldman, Valence focus and arousal focus: Individual differences in the structure of affective experience., Journal of personality and social psychology 69 (1995) 153.
[57] J. Ocumpaugh, Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual, New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences 60 (2015).
[58] M. S. Hussain, O. AlZoubi, R. A. Calvo, S. K. D'Mello, Affect detection from multichannel physiology during learning sessions with autotutor, in: Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15, Springer, 2011, pp. 131–138.
[59] E. Kannegieser, D. Atorf, J. Herold, Measuring flow, immersion and arousal/valence for application in adaptive learning systems, in: Adaptive Instructional Systems. Adaptation Strategies and Methods: Third International Conference, AIS 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II, Springer, 2021, pp. 62–78.
[60] N. Iddamalgoda, P. Thrimavithana, H. Fernando, T. Ratnayake, Y. Priyadarshana, R. Aththidiye, D. Kasthurirathna, A user-oriented ensemble method for multi-modal emotion recognition (????).
[61] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[62] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.
[63] P. Sarkar, A. Etemad, Self-supervised learning for ecg-based emotion recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3217–3221.

[64] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, H. Aronowitz, Speech emotion recognition using self-supervised features, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6922–6926.

[65] Y. Wu, M. Daoudi, A. Amad, Transformer-based self-supervised multimodal representation learning for wearable emotion recognition, IEEE Transactions on Affective Computing (2023).

[66] A. Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, I. Molenaar, Detecting emotions through electrodermal activity in learning contexts: A systematic review, Sensors 21 (2021) 7869.

[67] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125. URL: https://www.sciencedirect.com/science/article/pii/S1566253517300738. doi:https://doi.org/10.1016/j.inffus.2017.02.003.

[68] S. M. H. Mousavi, M. Besenzoni, A. Davide, A. Peternier, S. Giordano, The magic xroom: A flexible vr platform for controlled emotion elicitation and recognition, in: MobileHCI 2023 Demo Interactivity - Under Submission, ACM, 2023.

[69] R. Brinzea, B. Khaertdinov, S. Asteriadis, Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 01–08.

[70] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[71] S. Pramanick, L. Jing, S. Nag, J. Zhu, H. Shah, Y. LeCun, R. Chellappa, Volta: Vision-language transformer with weakly-supervised local-feature alignment, arXiv preprint arXiv:2210.04135 (2022).

[72] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, I. Misra, Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.

[73] A. Khare, S. Parthasarathy, S. Sundaram, Self-supervised learning with cross-modal transformers for emotion recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 381–388.

[74] K. Yang, T. Zhang, S. Ananiadou, Disentangled variational autoencoder for emotion recognition in conversations, IEEE Transactions on Affective Computing (2023) 1–12. doi:10.1109/TAFFC.2023.3280038.