Exploring Retrospective Annotation in Long-videos for Emotion Recognition

Bota, Patricia; Cesar, Pablo; Fred, Ana; da Silva, Hugo Placido

DOI
10.1109/TAFFC.2024.3359706

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Exploring Retrospective Annotation in Long-Videos for Emotion Recognition

Patrícia Bota, *Student Member, IEEE*, Pablo Cesar, *Senior Member, IEEE*, Ana Fred, *Member, IEEE*, and Hugo Plácido da Silva, *Senior Member, IEEE*

***Abstract*—Emotion recognition systems are typically trained to classify a given psychophysiological state into emotion categories. Current platforms for emotion ground-truth collection show limitations for real-world scenarios of long-duration content (e.g. >10 minutes), namely: 1) Real-time annotation tools are distracting and become exhausting; 2) Perform retrospective annotation of the whole content in bulk (providing highly coarse annotations); or 3) Are used by external experts (depending on the number of annotators and their subjective experience). We explore a novel approach, the EmotiphAI Annotator, that allows undisturbed content visualisation and simplifies the annotation process by using segmentation algorithms that select brief clips for emotional annotation retrospectively. We compare three methods for content segmentation based on physiological data (Electrodermal Activity (EDA), emotion-based), scene (time-based), and random (control) selection. The EmotiphAI Annotator attained a B+ System Usability Scale score and low-average mental workload as per the NASA Task Load Index (40%). The reliability of the self-report was analysed by the inter-rater agreement (STD < 0.75), coherence across time segmentation methods (STD < 0.17), comparison against the state-of-the-art ground truth (STD < 0.7), and correlation to EDA (>0.3 to 0.8), where the EDA-based method obtained the overall best performance.**

***Index Terms*—Annotation, emotion recognition, physiological signals, retrospective.**

I. Introduction

THE field of emotion recognition spans many applications, including wellbeing [1], e-learning [2], or recommendation systems [3]. The creation of emotion recognition systems requires the acquisition of large amounts of data for the development of artificial intelligence algorithms [4].

Manuscript received 6 March 2022; revised 6 November 2023; accepted 15 January 2024. Date of publication 29 January 2024; date of current version 2 September 2024. This work was funded by Fundação para a Ciência e a Tecnologia under Grant 2020.06675.BD, and in part by the Scientific Employment Stimulus-Individual under Grants Call-2022.04901.CEECIND/CP1716/CT0004, and PCIF/SSO/0163/2019 (SafeFire), and in part by the Scientific Employment Stimulus-Individual under Grant Call-2022.04901.CEECIND/CP1716/CT0004, and IT. Recommended for acceptance by G. McKeown. *(Corresponding author: Patrícia Bota.)*

Patrícia Bota, Ana Fred, and Hugo Plácido da Silva are with Instituto Telecomunicações and Instituto Superior Técnico, 1049-001 Lisbon, Portugal (e-mail: patricia.bota@tecnico.ulisboa.pt; afred@lx.it.pt; hsilva@lx.it.pt).

Pablo Cesar is with Centrum Wiskunde and Informatica, 1098, XG Amsterdam, The Netherlands, and also with the TU Delft, 2628, CD Delft, The Netherlands (e-mail: p.s.cesar@cwi.nl).

Digital Object Identifier 10.1109/TAFFC.2024.3359706

Peripheral psychophysiological signals have gained interest over the past years, as they provide a non-intrusive way of obtaining indicators related to the subjects' internal emotional state [4]. Sources such as the Electrodermal Activity (EDA) have shown to be correlated with the Sympathetic Nervous System (SNS), and have the advantage of enabling the integration into wearable systems that can be used to collect data in the wild.

In addition to the source input data, ground-truth collection is of paramount importance for the development of automated emotion recognition systems. The ground truth generally consists of the users' emotional states self-report and is used to train the algorithms and test their accuracy.

Current emotion annotation platforms are mostly desktop-based [5], [6] or perform the annotation in real-time [7], [8]. Real-time annotation can be distracting, as the user is required to annotate and visualise the content at the same time, thus it is usually limited to short videos so the annotation process doesn't become exhausting [4].

Previous work [9] reports that naturalistic (i.e. closer to real life) emotional physiological responses have shown to differ from lab-induced, with in-the-wild data collection showing higher accuracy. Additionally, in in-the-wild scenarios, large amounts of data can be collected for long periods and with higher frequency compared to doing appointments in a research facility with each subject.

For these reasons, the field of emotion recognition has been converging to data collection in the real world [9], [10]. We hypothesise that to collect in-the-wild emotion ground truth, an annotation platform should: 1) Allow the annotation of content experienced regularly in daily life, i.e. longer duration content (>10 minutes); 2) Allow an undisturbed emotional experience to maximise its elicitation power and be less intrusive; and 3) Simplify the emotion annotation process to enable frequent annotations by the subjects and the collection of large amounts of data. Current annotation platforms either perform a single post-hoc annotation (after the content visualisation) or perform the annotation while watching (real-time). These approaches do not meet our hypotheses since the unique post-hoc annotation over-generalisation inherently loses information, and real-time annotation platforms do not allow an undisturbed visualisation of the content, with users becoming exhausted in longer-length annotations.

As a result, we extend the state of the art (SoA), by exploring a novel approach based on retrospective annotation of long-duration content – the EmotiphAI Annotation tool. As

TABLE I
PREDOMINANT DATASETS FOR EMOTION RECOGNITION WITH PHYSIOLOGICAL DATA

Name	#S	Stimuli	Loc	Area	Annotation	Sensor
DEAP [17]	32	40 music videos (1min)	L	**ER**	A, V, D, liking, familiarity	ECG, **EDA**, EEG, Electromyography (EMG), EOG, Respiration (RESP), Skin Temperature (SKT), face video
MAHNOB [18]	27	20 film clips (35-117s)	L	**ER**, implicit tagging	A, V, D, predictability, BET	ECG, **EDA** EEG, RESP, SKT, face and body video, eye gaze tracker, audio
ASCERTAIN [19]	58	36 movie clips (51-128s)	L	**ER**, Implicit personality	A, V, liking, engagement, familiarity, Big Five	ECG, **EDA**, EEG, facial video
Eight-Emotion [20]	1	8 posed emotions	L	**ER**	Neutral, anger, hate, grief, joy, platonic love, romantic love, reverence	ECG, **EDA**, EMG, RESP
EMDB [21]	32	52 non-auditory film clips (40s)	L	**ER** w/out auditory content	A, V, D	**EDA**, HR
AMIGOS [12]	40	16 250s videos; 4 videos (14min) alone and in group	L	**ER**, personality traits and mood on individuals and groups	Big-Five personality traits, PANAS, A, V, D, liking, familiarity and BET	Audio, visual, depth, EEG, GSR and ECG
DECAF [22]	30	40 music video (1min DEAP), 36 movie clips	L	**ER**	A, V, D	ECG, EMG, EOG, MEG, near-infrared face video
WESAD [23]	15	baseline (20mins), video clips (392s), TSST (5min), meditation (7mins)	L	**ER**, SR	SAM, PANAS, SSSQ and STAI	ACC, ECG, **EDA**, EMG, RESP, SKT; ACC, **EDA**, **PPG**, SKT
CASE [24]	30	Video clips (101-197s)	L	**ER**	A, V	ECG, **Photoplethysmography (PPG)**, EMG (3x), **EDA**, RESP, SKT
CLAS [25]	62	Math and logic problems, Stroop test, IAPS, multimedia clips (30min)	L	**ER**, SR	A, V	ECG, **PPG**, **EDA**, ACC
PAFEW [26]	24	480-6040ms short videos	L	**ER**	7 BET	**EDA**, **PPG**, SKT, pupil
K-EmoCon [24]	32	Debate (10min)	**WL**	**ER**	Retrospective A, V, 18 BET	EEG, ECG, **PPG**, **EDA**, SKT
COGNIMUSE [13]	7	Movies	**WL**	**ER**	Feeltrace A, V	None
LIRIS-ACCEDE [14]	10	30 movies	**WL**	**ER**	GTrace A, V	**EDA**, motion, SKT

Number of subjects (#S); setting (LOC): lab (L), in-the-wild (W); application: emotion recognition (ER), stress recognition (SR); annotation space: arousal (A), valence (V) and dominance (D), basic emotions theory (BET). Table adapted from [16].

an alternative to real-time annotation of the entire content, we propose the use of an intelligent segmentation algorithm that selects brief clips (10 seconds) based on pre-defined criteria for emotional annotation retrospectively. We test three content segmentation approaches: 1) EDA-based – SNS-based segmentation taking into consideration that EDA is a marker of SNS activity (in particular Arousal) [11]; 2) Scene-based – time-based using scene boundary detection algorithms described in the SoA (i.e. PySceneDetect); and 3) Random selection – used as a null hypothesis. Lastly, the EmotiphAI Annotator is validated for its usability and reliability across a comprehensive set of metrics.

II. RELATED WORK

A. Datasets

There are multiple publicly available datasets for emotion recognition based on physiological data. They allow researchers to develop and compare different algorithms on the same data, without the experimental effort of collecting it. Table I summarizes prevalent datasets in the emotion recognition SoA, with their main characteristics.

Most datasets for emotion recognition acquire data in controlled, in-lab scenarios. The use of small video clips is predominant, except for three datasets that although containing data collected in laboratory settings, are closer to naturalist scenarios, i.e. by using long-duration videos (>10 minutes) as the elicitation method, namely: 1) AMIGOS [12]; 2) COGNIMUSE [13]; and 3) Continuous LIRIS-ACCEDE [14]. The AMIGOS [12] dataset contains data from 16 short emotional videos, and 4 long videos (alone and in group settings) with ≈ 14 to 24 minutes, including EDA, Electrocardiography (ECG), and Electroencephalography (EEG) data. The long video data was segmented in 20-second windows, and annotated using the individuals' facial expressions by three external annotators in the continuous arousal and valence dimensions using the Self-Assessment Manikin (SAM). The COGNIMUSE [13] dataset contains data from 7 Hollywood movies annotated in arousal and valence. Two types of annotations are provided: experienced self-reports by the volunteers; and, intended emotions annotated by experts. Both types of annotation are performed using the FEELTRACE platform [5], and no physiological data is acquired. Continuous LIRIS-ACCEDE [14] contains data from 30 movies (M = 884 s, STD = 766 s) annotated on valence or arousal axes by a modified GTrace platform [15] incorporating SAM and a joystick. During video visualisation, EDA, finger motion, and skin temperature were recorded.

Discussion: The SoA shows that, overall, the available datasets for emotion recognition based on physiological data, using videos as the emotion elicitation method, rely on small video clips (excerpts extracted from longer videos) as the elicitation method. The use of clips lasting a few seconds to a few minutes detaches the elicitation from a naturalistic elicitation, where longer-duration content such as the entire movie/TV show is watched. Media content is usually devised taking into account an emotional timeline that builds to the elicitation of a certain emotion, such that the use of small video clips can reduce the elicitation power of the content.

We address this issue by proposing a novel methodology for the annotation of longer videos. With the usage of longer videos, our goal is to approximate the elicitation process to a naturalistic scenario for the creation of datasets representative of the real world. For the same reason, we target an undisturbed

TABLE II
PREDOMINANT EMOTION ANNOTATION PLATFORMS IN THE EMOTION RECOGNITION SOA

Platform	Device	Model	Rank	RT	Cont.
FeelTrace [5]	Desktop	V, A	✗	✓	✓
Gtrace [15]		1d	✗	✓	✓
DARMA [28]		2d	✗	✓	✓
PAGAN [29]		1d	✓/✗	✓	✓/✗
AffectRank [6]		V, A	✓	✓	✗
AffectButton [30]		Image	✗	✗	✗
RankTrace [27]		1d	✓	✓	✓
NOVA [31]		V,A, Tags	✓	✗	✓
RCEA [7]		V, A	✗	✓	✓
EmoWheel [32]	Mobile	Tags	✗	✓	✗
EmoteU [33]		V, A	✓/✗	✓	✓
EmotiphAI	Web	V, A	✗	✗	✗

RT (real-time); cont. (continuous).

visualisation of the content with retrospective annotation as an alternative to real-time annotation.

B. Annotation Platforms

Table II shows predominant emotion annotation platforms in the SoA, analysed by their main characteristics. The SoA platforms are divided into mobile and desktop-based. Mobile applications allow out-of-the-lab use and have a smaller form factor. Desktop applications, often relying on additional peripherals such as joysticks [8] or wheel mouses [27], are predominant.

Most platforms rely on the arousal and valence dimensions, either through Russel's Circumplex model [6], [7], or Lang & Bradley's SAM manikins [17], [34], [35]. One exception is RankTrace [27], in which volunteers rate their tension using a wheel mouse while watching a video. When performing real-time annotation, one dilemma is the use of one or two dimensions. In [27], [29], only one dimension at the choice of the researcher is used, while in most works [5], [6], [7] the 2D circumplex model is used. While one dimension might be insufficient to obtain a comprehensive description of emotion, the use of two dimensions increases the annotation mental workload and is distracting, limiting engagement in the elicitation content.

While traditional platforms annotate a magnitude value, rank-based platforms [6], [27], [29] annotate the emotion relative change. In AffectRank [6], the annotation scale is left unbounded, with the results showing higher inter-rater agreements when compared to a bounded annotation.

The emotion annotation can be performed in real-time while the subject experiences the emotion, or post hoc, after the content, retrospectively. Continuous annotation is inherently conducted in real-time, enabling the capture of the temporal dynamics of emotions. Discrete methods such as SAM or questionnaires applying basic emotion theory are usually annotated in post-hoc [17], [34], [35].

Both real-time and post-hoc annotation show similar mental workload values in [7], [33], while lower values for real-time were reported in [36]. Continuous emotion models are predominant [5], [7], [29].

In opposition to traditional annotation, the NOVA [31] platform interactively incorporates semi-supervised algorithms to pre-label data automatically. The semi-supervised annotation is then manually annotated using a user-friendly interface that displays confidence levels and visual explanations to streamline the annotation process.

TABLE III
NUMBER OF INDIVIDUALS PER MOVIE, AGE AND GENDER DISTRIBUTION.

Movie	Total Ind. (#)	Age (years)	Gender (%)
After the Rain	18	21.4 ± 1.9	F: 37, M: 63
Elephant's Dream	12	21.5 ± 1.6	F: 32, M: 68
Tears of Steel	15	21.9 ± 1.5	F: 21, M: 79

Discussion: The SoA (Table IV) shows that the majority of the platforms perform continuous real-time annotation or a single post-hoc discrete annotation of the entire elicitation process. Both are useful to annotate small video clips (<10 minutes). However, these platforms are not fit to annotate real-life emotional experiences, usually of longer duration, e.g. TV show episodes (≈25 to 45 minutes), films (≈2 hours), or theatre performances (>1 h). In longer-duration elicitation methods, real-time annotation becomes distracting and exhausting, and the use of a single post-hoc discrete annotation does not capture detailed information about emotional events. This reinforces the motivation for the annotation of long videos. EmotiphAI Annotator aims to be a halfway house between the two approaches (real-time and retrospective), with a stepped retrospective approach allowing an undisturbed emotion elicitation and performing retrospective annotation of selected emotional events.

C. Processes Involved in Emotional Self-Report

Self-reports can be acquired anywhere, quickly and regarding any event (e.g., about how the subject felt in the past, how will they feel in the future, or even in hypothetical situations). There is a long tradition of offering retrospective annotations over days, weeks, months [37], after 90 days [38] or even after 1 a [39], without losing their validity and reliability. What changes between different types of annotation are the inherent processes involved in the subject's sources for the self-report. When the current emotion is not accessible, subjects resort to their memory or their identity-related beliefs and values.

The authors in [37] describe that real-time annotations tend to rely on direct feelings (experiential knowledge), while for retrospective annotations episodic memory is used. Although past emotional events can not be re-experienced, they can be reconstructed by anchoring on relevant thoughts, events or thoughts. With this in mind, in the EmotiphAI Annotator, the re-visualization of the video is used to provide context as an anchoring tool to help the reconstruction of the emotion. This is also corroborated by [40], where the volunteer's request to improve annotation would be to have context regarding the annotation. It is also supported by the work of [41], where contextual information improved the inter-rater agreement.

The episodic memory and the ability of the subject to recall past emotions, like any memory, decreases with time. The subject will then use situation-specific and identity-related beliefs. The two become relevant when asking for hypothetical, prospective and trait-related emotions. In [42], the authors denoted

TABLE IV
ANNOTATION STATISTICS: AVERAGE ANNOTATION TIME BETWEEN CLIPS (*ANN. TIME (S)*), NUMBER OF ANNOTATIONS PER SUBJECT (*NUM. ANN.*), TOTAL NUMBER OF ANNOTATIONS (*TOTAL ANN.*), AVERAGE (*ANN. AVG TS.*), MAXIMUM (*MAX ANN. TS.*) AND MINIMUM (*MIN. ANN. TS.*) NUMBER OF ANNOTATIONS PER TIMESTAMP

Dimension	Method	Ann. Time (s)	Num. Ann. (#)	Total Ann. (#)	Ann. Avg Ts. (#)	Max Ann. Ts. (#)	Min. Ann. Ts. (#)
Arousal	Scene	25.98 ± 0.37	15.41 ± 0.98	468.00 ± 144.87	14.65 ± 2.75	15.00 ± 2.45	14.00 ± 2.45
	Random	27.55 ± 1.02	10.19 ± 0.59	322.33 ± 132.32	09.93 ± 1.43	13.33 ± 1.70	01.00 ± 0.00
	EDA	**25.14 ± 0.06**	12.84 ± 0.25	400.67 ± 148.87	12.44 ± 2.12	**15.00 ± 2.45**	01.00 ± 0.00
Valence	Scene	25.63 ± 0.62	15.43 ± 1.00	468.33 ± 144.49	14.65 ± 2.79	15.00 ± 2.45	14.00 ± 2.45
	Random	27.40 ± 0.98	10.17 ± 0.58	321.67 ± 131.94	09.89 ± 1.48	13.33 ± 1.70	01.00 ± 0.00
	EDA	**24.90 ± 0.11**	12.82 ± 0.24	400.00 ± 148.48	12.43 ± 2.14	**15.00 ± 2.45**	01.00 ± 0.00

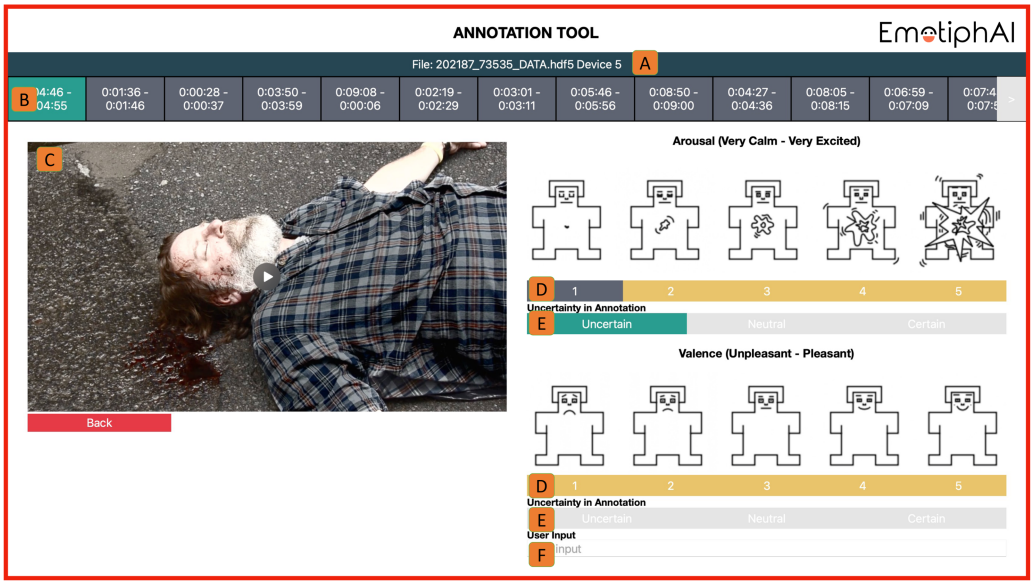

Fig. 1. EmotiphAI Annotator self-reporting annotation interface.

that daily tiredness and big-five personality traits [43] influence retrospective annotations over a 1-day and 2-week period. Individuals who score high in neuroticism often recall more negative emotions compared to what they report in real-time assessments, akin to how extroverts typically remember more positive emotions [38].

Bias is present in all types of emotional self-reports. Age has been shown to influence both momentary and retrospective annotation [44], with older adults being optimistic and rating emotions more positively [39], [45]. Social status can also hinder a true report [46], [47], as well as psychological issues such as alexithymia (the inability to describe one's emotional states) [47], [48].

III. PROPOSED APPROACH

We introduce a retrospective annotation tool for emotion assessment in longer videos – the EmotiphAI Annotator. It allows the annotation of longer-duration content by using an intelligent algorithm that simplifies the annotation process by selecting small segments (10 seconds) for the user to annotate retrospectively, instead of annotating the entire content. In contrast to the existing platforms, by performing the annotation retrospectively, EmotiphAI allows the subject to fully enjoy and be engaged in the content, thus, not increasing the mental workload during visualisation by annotating. EmotiphAI consists of a web-based application that can be used both on mobile and desktop (working across all operating systems) and requires no additional material (e.g. computer mice or joystick).

A. Annotation Interface

Fig. 1 shows the EmotiphAI Annotator user interface for emotion assessment. The users are anonymously identified by their device ID at the top of the page (A) in a dropdown menu. Below, a series of 10-second segments are given for the user to annotate (B). When the page is loaded, the first segment is automatically selected. In the video player on the left (C), the user can press play and review the video segment. Through visualisation of the content media, the user should be able to recall their emotional status during the initial visualisation of the clip and report their retrospective emotional state. For the annotation questionnaire, we rely on a validated emotion scale, the SAM [35] arousal – valence (D), aided by graphics of manikins to express emotional states. Several factors may introduce uncertainty in the annotation, namely, the annotator may not fully understand the concept of valence or arousal, the annotator may not be engaged in the study and respond randomly, and the segment may contain several emotions, among others. For this reason, we introduced uncertainty (E) self-reporting, which enables the user to detail their level of confidence in the annotation. Lastly, an optional user input box (F) is given for providing additional comments,

e.g. indicate an external interference. Once all the information is reported, the next segment will automatically load for the user to annotate.

The EmotiphAI back-end was written in Python 3, and the front-end user interface in HTML, CSS and JavaScript, with the communication between the two performed using Flask.[1] A Raspberry Pi was used as a set-top box to run the EmotiphAI Annotator software. For further information regarding the EmotiphAI data acquisition infrastructure, we refer the reader to [49].

B. Content Segmentation Method

The platform contains integrated algorithms that segment moments of the film for emotion assessment:

EDA: Selects for annotation, the moments of the video where the subjects' EDA data presented an emotional onset, considered as a minimum on the EDA data followed by an increase of the data above a threshold of 0.01% of the EDA maximum [50]. The annotation segment starting time was marked to be 4 seconds before the EDA onset, to take into consideration the latency of emotional stimuli (between 1 to 5 seconds [51]). The data was pre-processed by the application of a low pass 1 Hz cut-off frequency and average window (20 seconds) to remove high-frequency noise and only detect relevant changes in the EDA.

The segments were ordered by their amplitude following the literature that reports that *"Our most vivid memories tend to be emotional"* [52]. Higher emotional events are easier to recall and, thus, are shown first, ensuring that the annotation of the high-intensity EDA events is made (in a real-world scenario subjects may not complete the rating of all the segments). The EDA-based approach was developed to identify high-intensity SNS-related events. Moreover, EDA fluctuations can occur due to various reasons, including temperature changes, movement, or external distractions. By focusing on higher amplitude EDA event, we aim to reduce the chances of annotating segments that might be influenced by these external factors. As manual annotation is a resource-intensive process, focusing on segments with high amplitude EDA allows us to make the most out of our annotation efforts.

Scene: Selects for annotation the last 10 seconds of every movie scene so a time-sequential annotation is performed. The scenes were identified by existing SoA methods for scene boundary detection based on image features, namely the PySceneDetect[2] content-aware scene detector, which cuts scenes where the difference between frames exceeds a pre-defined threshold (of 60 as given by [53]), empirically found to detect both abrupt transitions and fades to black. The decision to annotate only the last 10 seconds of every scene was taken in order to capture the most emotional dynamics from that particular scene. We assume that the emotional intensity or resolution would be most pronounced towards the end, offering the most representative snapshot for the annotation.

Random: Used for control, it randomly selects 10-second clips from the entire video. Both the number of segments for annotation and their instant are randomly selected using a discrete uniform distribution, taking into consideration the length of the video clip. The number of segments is obtained by a random selection from $\sqrt{l}/2.5$ to $\sqrt{l}/1.5$, being l the length of the movie. The random selection method was introduced to compare the obtained results from the aforementioned techniques to random chance. That is, to see if there is a preferred method to extract meaningful clips for annotation in a long-duration film or, on the other hand, if the entire film is equally meaningful for emotional ground-truth collection.

[1] https://flask.palletsprojects.com/en/1.1.x/

[2] https://pyscenedetect.readthedocs.io/en/latest/

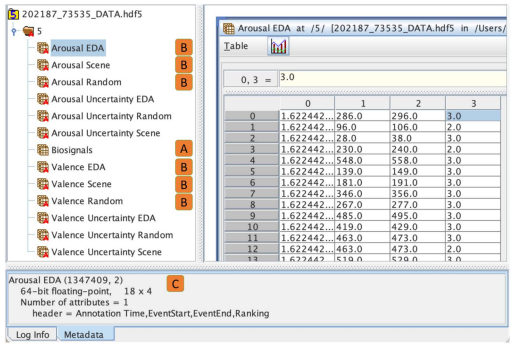

Fig. 2. HDF5 file storing the physiological data, user's annotations and video information. Column information can be found in [C] – header.

C. Data Storage

The user's annotations are stored in an HDF5 file. Fig. 2 illustrates an example file generated by the EmotiphAI Annotator. The data is stored in a hierarchical structure: For each user (e.g. user with device ID 5) a dataset is created, storing both the physiological data (i.e. Fig. 2-A) and the user's annotations for the segmentation algorithms (Fig. 2-B). For each segmentation algorithm and emotion dimension (arousal, valence and uncertainty), the data is gathered in a group. On the bottom (Fig. 2-C), the information regarding the metadata for each group is shown, with the number of annotations and a description of the data stored in each column (header attribute). For each annotation the EmotiphAI Annotator stores: annotation time (in seconds), annotated segment start time (in seconds), annotated segment end time (in seconds), and self-report value (ranking).

IV. Experimental Study Design

To analyse the EmotiphAI Annotator usability and annotation reliability, an experimental study was performed. The study was submitted to and approved by the Instituto Superior TÃ©cnico - University of Lisbon Ethics Committee (Ref. n. ° 11/2021 Date: 20/04/2021).

Procedure: To ensure the correct deployment of the protocol, an assistant was present during the session and guided the subjects across the steps shown in Fig. 3. Steps 6 and 7 were repeated three times by every user, one time for each segmentation method.

Video Stimuli: Three videos were extracted from the Continuous LIRIS-ACCEDE dataset [14]. The videos were selected

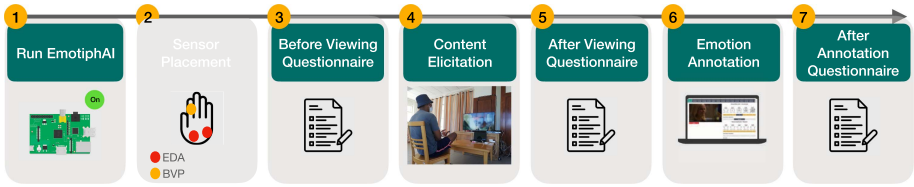

Fig. 3. Experimental protocol.

taking: 1) Length of the movie – Videos were limited to medium-length videos (i.e. M = 10.72 minutes, STD = 1.09 minutes) to reduce the time and complexity of the protocol; and 2) Valence-Arousal Space – Movies eliciting the four quadrant areas of the valence-arousal space; **Participants:** 19 participants were recruited to watch 3 videos. The volunteers' gender and age distribution is shown in Table III. The volunteers varied across videos according to their availability. No participant reported any concerning health condition that would impact the study.

Physiological Measures: During the video visualisation BITalino [54] EDA and PulseSensor PPG[3] data were acquired with accelerometer, gyroscope, magnetometer, and attitude (pitch/yaw/roll) data. The data was collected using an R-IoT microcontroller [49]; with pre-gelled self-adhesive Ag/AgCl electrodes connected on the non-dominant hand hypothenar and thenar eminences. Data acquisition was performed at a 60 Hz sampling rate with 12-bit resolution.

Table III shows the number of individuals, ages and genders per video after individuals with "erroneous" acquisitions were removed. We considered an "erroneous" acquisition to happen when the data was at 0 for long periods, or when there was a problem in the annotation process and the volunteers did not annotate across the three segmentation algorithms.

V. Results

The EmotiphAI Annotator's usability and reliability were evaluated across a set of comprehensive tests. Each self-report annotation was extended to a size of 60 seconds. To compare the segmentation methods, the video was segmented at the physiological data sampling period, with the sample points considered as a timestamp. This designation facilitated the temporal alignment and comparison of the annotation method outputs at specific time points. Only timestamps with annotated segments are used in the comparison analysis, as only these have values to be compared for each method. We posit that no segments are incorrectly selected, as in the optimal case we would have the entire movie annotated; however, as it was previously discussed, that is impractical due to its high cost for broad annotation of long-term content.

Four validity tests were performed to analyse if the EmotiphAI Annotator's retrospective annotations are accurate and reliable.

A. Usability Test

Table IV presents a summary of the annotations obtained with the content segmentation methods. The results show that, although the number of segments for annotation was lower for the Random algorithm (*Num. Ann.* column), the volunteers took on average a lower time annotating the segments identified using the EDA method (*Ann Time (s)* column). The *Total Ann.* column shows that the Scene algorithm identified the largest number of clips for annotation, while the Random algorithm identified the lowest number. The columns *Ann. Avg Ts.*, *Max Ann. Ts.*, and *Min. Ann. Ts.* show the average, maximum and minimum number of coinciding annotations performed by the subjects in a given timestamp. A coinciding annotation is observed if there is an annotation in a timestamp for more than one user. The Scene detection algorithm has the largest number of coinciding annotations on average (*Ann. Avg Ts.*), followed by the EDA. Although the EDA detects a lower number of segments for annotation than the Scene method, the maximum number of coinciding annotations in a timestamp is equal. This indicates that the EDA selected at least one common region for annotation across all the subjects, i.e. one section of the movie had a high impact on all the subjects, which was reflected in the EDA data. The minimum number of annotations (*Min. Ann. Ts.*) shows that there were regions in which only one annotation (by one subject) was selected using the EDA and Random segmentation. These numbers are expected as the scene detection will select the same segments for every user, the random segmentation should select uncorrelated random segments, and the EDA should lead to segments that can be similar (corresponding to high arousal). The different number of annotations across subjects is observed when the subjects forget or skip a clip for annotation.

TABLE V
EmotiphAI Annotator Usability Questionnaires

Method	SUS (%)	NASA-TLX (%)	Preference (%)
Scene	73.50 ± 1.89	**44.74 ± 0.58**	17.67 ± 7.64
Random	74.75 ± 3.87	45.21 ± 1.30	**51.00 ± 3.46**
EDA	**76.40 ± 2.11**	45.56 ± 1.48	31.33 ± 5.51

The platform usability and user experience assessment was performed using the System Usability Scale (SUS) [55], and the mental workload was evaluated using the NASA Task Load Index (NASA-TLX) [56]. A preference question was also introduced. The results are shown in Table V. The SUS results show a preference for the EDA method with a B^+ grade [57]. While for the NASA-TLX, similar values are reported across the three algorithms ($\approx 45\%$). The lower value is obtained for the Scene algorithm and the larger for the EDA method. The obtained values report that EmotiphAI is in the top 40% of the platforms tested by the NASA-TLX questionnaire, corresponding to an average-to-low mental workload. Lastly, the preference question shows that the subjects tended to select as their favourite method the algorithm with a lower number of segments for annotation, i.e. the Random method, followed by the EDA and Scene methods.

[3]https://pulsesensor.com/

TABLE VI
EMOTIPHAI ANNOTATOR INTER-SUBJECTS AGREEMENT RESULTS, GIVEN BY THE EVALUATION ERROR (*EVAL. ERROR*) AND *STANDARD DEVIATION (STD)* IN A GIVEN TIMESTAMP

D	Method	Eval. Error	STD	d Opt.	d Max.
A	Movie	0.18 ± 0.12	0.69 ± 0.13	0.51 ± 0.0	2.05 ± 0.01
	Random	0.19 ± 0.18	0.71 ± 0.18	0.52 ± 0.0	2.08 ± 0.02
	EDA	0.22 ± 0.17	0.74 ± 0.18	0.52 ± 0.0	2.07 ± 0.01
V	Movie	0.05 ± 0.05	0.55 ± 0.07	0.51 ± 0.0	2.05 ± 0.01
	Random	0.05 ± 0.07	0.54 ± 0.09	0.52 ± 0.0	2.08 ± 0.02
	EDA	0.04 ± 0.06	0.55 ± 0.08	0.52 ± 0.0	2.07 ± 0.01

The optimal (d Opt.) and maximum upper bound (d Max.) rate the results performance. Annotations $\in \{1, 5\}$. D (dimension); a (arousal); v (valence).

B. Inter-Subject Agreement:

A common approach in the SoA to validate self-reports is to calculate the subject's inter-rater agreement, e.g. standard deviation and evaluation error in [58], correlation, Krippendoff's α ordinal, Cohen's k, difference to mean, mean absolute difference in [59], Krippendorff's α in [6], [29], and Fleiss' Kappa in [36]. In this work, we use the metrics applied in [58], where, similarly as in this work, naturalistic data was collected and the annotations were performed using the SAM. In [58], the authors obtain the inter-subject agreement by analysing the annotation's standard deviation in a given timestamp, and the evaluation error given by the difference between the average standard deviation in a timestamp and the optimal standard deviation bound [58]. The difference was defined as zero when the standard deviation was below the optimal bound. The optimal and maximum thresholds were obtained using (1) & 2 [58], respectively, where K is the number of annotators.

$$d_{opt} \leq \frac{1}{2}\sqrt{\frac{\xi}{\xi - 1}}, \text{ with } \xi = \begin{cases} K, & K \text{ even} \\ K + 1, & K \text{ odd} \end{cases} \quad (1)$$

$$d_{\max} \leq 2\sqrt{\frac{\xi}{\xi - 1}}, \text{ with } \xi = \begin{cases} K, & K \text{ even} \\ K + 1, & K \text{ odd} \end{cases} \quad (2)$$

The experimental results for the EmotiphAI Annotator are shown in Table VI. Similar results are obtained for the three segmentation algorithms, with a lower evaluation error obtained for the valence dimension. For both dimensions, the standard deviation is above the optimal upper bound and below the maximum bound. The evaluation error values are lower than half the distances between two discrete SAM numbers, being very low for the valence dimension, and the STD between one range value.

C. Self-Reports Coherence

The self-report coherence was analysed by observing if the three algorithms lead to similar time-series annotations for each movie. The evaluation is obtained by comparing the user's average annotations for each timestamp across the three algorithms using the metrics from [58] detailed in Section V-B (Table VII). The experimental results show an evaluation error of 0. The zero value is obtained due to the standard deviation being below the optimum bound, returning an evaluation error of 0. This leads to the conclusion that, for the three segmentation algorithms, the average annotations are similar across timestamps. The valence dimension shows lower variability. An example of the subjects' average annotations for the "Elephant's Dream" movie is shown in Fig. 4. The figure confirms that there is negligible inter-segmentation method variability, with an overall value below a magnitude of 1.

TABLE VII
SIMILARITY BETWEEN THE SUBJECTS' AVERAGE SELF-REPORTS FOR THE SEGMENTATION METHODS

Dimension	Eval. Error	STD	d Opt.	d Max.
Arousal	0.0 ± 0.0	0.17 ± 0.07	0.58 ± 0.0	2.31 ± 0.0
Valence	0.0 ± 0.0	0.14 ± 0.02	0.58 ± 0.0	2.31 ± 0.0

Annotations $\in \{1, 5\}$.

TABLE VIII
COMPARISON BETWEEN THE EMOTIPHAI ANNOTATOR AND THE LIRIS-ACCEDE ANNOTATIONS

D	Method	Eval. Error	STD	d Opt.	d Max.
A	Scene	0.06 ± 0.08	0.65 ± 0.19	0.71 ± 0.0	2.83 ± 0.0
	Random	0.04 ± 0.05	0.64 ± 0.17	0.71 ± 0.0	2.83 ± 0.0
	EDA	0.03 ± 0.04	0.56 ± 0.18	0.71 ± 0.0	2.83 ± 0.0
V	Scene	0.02 ± 0.03	0.49 ± 0.20	0.71 ± 0.0	2.83 ± 0.0
	Random	0.00 ± 0.00	0.62 ± 0.06	0.71 ± 0.0	2.83 ± 0.0
	EDA	0.05 ± 0.07	0.56 ± 0.21	0.71 ± 0.0	2.83 ± 0.0

Annotations $\in \{1, 5\}$.

D. Comparison to Reference Annotations

The annotations obtained using the EmotiphAI Annotator were compared to the annotations reported in LIRIS-ACCEDE [14], using the metrics proposed by the authors in [58], detailed in Section V-B (Table VIII). The LIRIS-ACCEDE baseline was used since it was the only publicly available corpus with physiological data and annotated longer videos in the valence and arousal dimensions identified by the authors. The experimental results show that, for each timestamp, the evaluation error is similar across dimensions and segmentation algorithms, with no segmentation method outperforming the remaining. A similar result for the different segmentation algorithms is expected, since the self-reports coherence results in Section V-C showed that the average annotations across the algorithms are similar. The obtained standard deviation is below the optimal value and the evaluation error is minimal. An example is shown in Fig. 5, with the comparison of the SoA ground-truth (red) with the obtained annotations using the EmotiphAI algorithms (EDA in green, Random in orange and Scene-based annotations in blue). Although a drift is observed between EmotiphAI's and LIRIS-ACCEDE annotations, on average, the standard deviation value between the two is low (below a magnitude of 0.7 in a [1, 5] scale). Fig. 5 also shows that the annotations time series from both datasets tend to follow the same trend pattern throughout the movie.

E. Comparison to Electrodermal Activity

Taking into consideration that EDA is a marker for SNS activity (arousal) [4], an analysis is performed as to whether there is a correlation between the EmotiphAI Annotator's arousal self-reports and the collected EDA data. To process the EDA data, we resort to the validated algorithm described in [60], where,

Fig. 4. Average arousal and valence annotations for the "Elephant's Dream" movie.

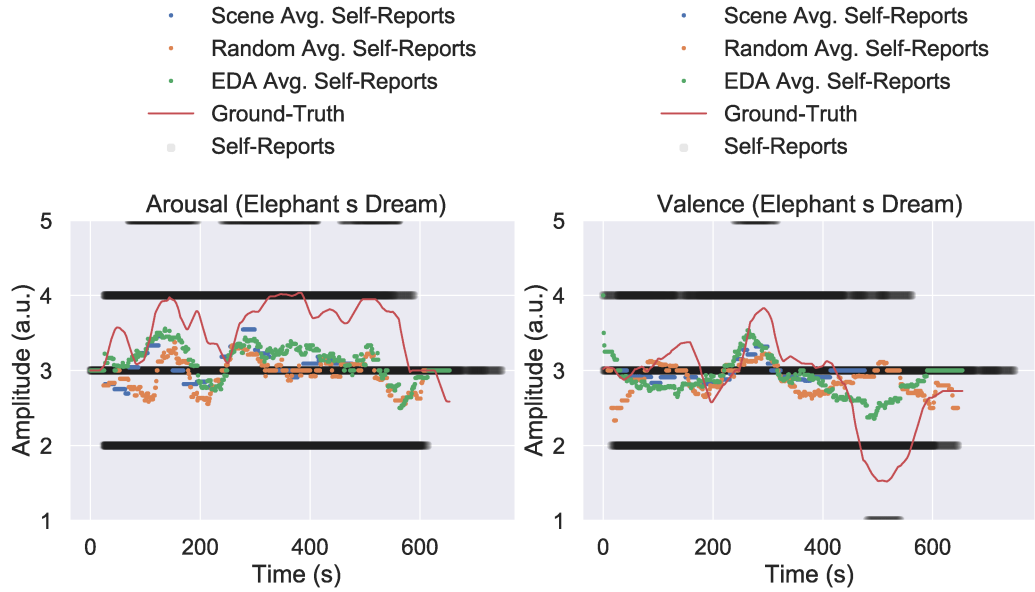

Fig. 5. Examples comparing the LIRIS-ACCEDE ground-truth to the EmotiphAI MAP segmentation methods. The annotated segments are shown in grey, becoming black when multiple annotations are superimposed.

similarly to this paper, the subjects' EDA given by the Mean Affective Profile (MAP) was compared to the LIRIS-ACCEDE ground-truth used as a benchmark. The MAP was validated in [61] to contain information regarding the arousal variations of a global audience during a movie. It involves the removal of outlier subjects, a low-pass filter, taking the first derivative, truncation of positive values, and downsampling through a moving average filter. For further information regarding the MAP, the reader is referred to [60], [61].

The experimental results in Table IX show that the segmentation correlation results are dependent on the movie. Nevertheless, comparing the three algorithms, and throughout the three movies, we observe that the EDA method is the only method maintaining a lower-average to above-average correlation between the EDA MAP and the arousal annotations. Poorer results are obtained for the video "Elephant's Dream", which may be explained by the low self-reports variability (M = 3.33, STD = 0.10). The data from the "Tears of Steel" and "After the Rain" movies is shown in Fig. 6, where an increase in the annotation value is followed by an increase in the EDA MAP, and vice-versa.

TABLE IX
EMOTIPHAI ANNOTATOR CORRELATION TO EDA GIVEN BY THE PEARSON (*PEARSON C*) AND SPEARMAN (*SPEARMAN C*) CORRELATION BETWEEN THE SUBJECTS' AVERAGE AROUSAL SELF-REPORTS AND THE SUBJECTS' WEIGHTED EDA PROFILE IN A [0, 1] SCALE

Movie	Dimension	Pearson C	Spearman C
Tears of Steel	Scene	**0.77**	**0.73**
	Random	0.58	0.65
	EDA	0.58	0.55
After the Rain	Scene	0.62	0.31
	Random	0.32	0.18
	EDA	**0.80**	**0.65**
Elephant's Dream	Scene	-0.10	-0.18
	Random	-0.21	-0.20
	EDA	**0.26**	**0.25**

In Table X the EmotiphAI Annotator's best results are evaluated comparatively to the results obtained by the SoA [60]. As it

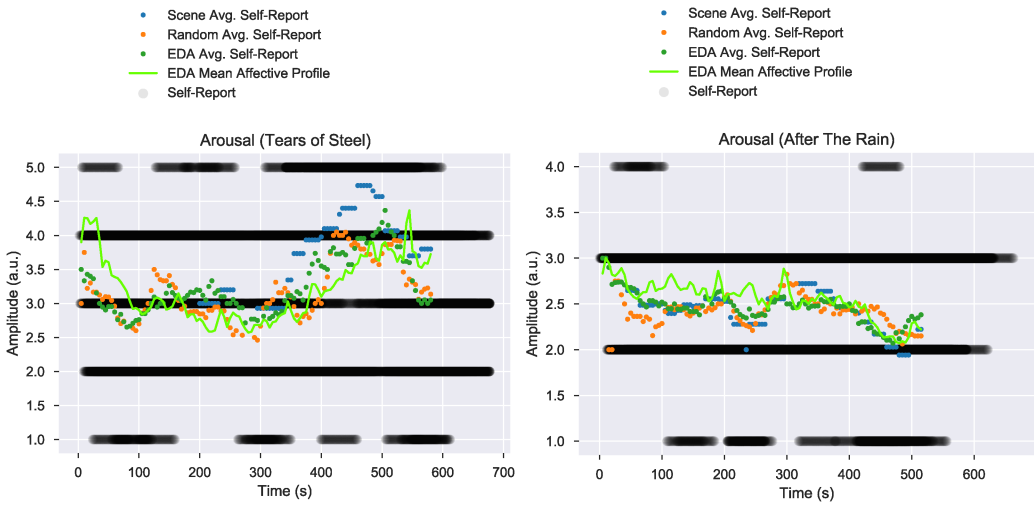

Fig. 6. Example comparing the collected EDA and the average arousal self-reports from the EmotiphAI Annotator.

TABLE X
COMPARISON BETWEEN THE SOA [60] AND EMOTIPHAI ANNOTATOR BEST RESULTS FOR THE PEARSON (*PEARC. C*) AND SPEARMAN (*SPEAR. C*) CORRELATION BETWEEN THE MAP AND THE AROUSAL ANNOTATIONS

Movie	Database	T (%)	Pearc.	Spear	Method
ToS	LA	66	0.55	0.64	
	EAI	60	**0.77**	**0.73**	Scene
AtR	LA	86	0.24	0.27	
	EAI	50	**0.80**	**0.65**	EDA
ED	LA	66	0.55	0.64	
	EAI	90	0.26	0.25	EDA

ToS ("tears of steel"); AtR ("after the rain"), ED ("elephant's dream"); LA (LIRIS-ACCEDE); EAI (EmotiphAI).

can be seen, EmotiphAI Annotator obtains a higher correlation between the arousal self-reports and the EDA MAP, outperforming the SoA annotations for two of the movies ("Tears of Steel" and "After the Rain"), where there was an observed variability in the EDA self-reports correlated to an emotional event.

VI. DISCUSSION

A. Usability Test

The results show that, although retrospective emotion annotation is a task that requires memory work and increases the subject mental workload, it is not overwhelming and too tiring for the volunteers as shown by the NASA-TLX of 40%. This value is in line with the SoA, namely with EmoteU [33] (37.5% to 44.52%), and RCEA [7] (52.5% and 82.5%) and above [36], where live (31.6%) and textual (35.7%) annotation is performed. Amongst our two hypotheses under test (EDA and Scene-based segmentation), the EDA segmentation was the preferred method. Although the EDA method selected a higher number of clips for annotation than the Random method, it showed the lowest annotation time of all the methods. The SUS B+ score confirms the usability of the EmotiphAI Annotator for retrospective annotation, with the EDA method outperforming the Scene-based in terms of annotation time, SUS and preference score.

B. Validity Test

The EmotiphAI Annotator reliability was analysed through a set of four tests:

Inter-subject Agreement: The evaluation error was lowest for the valence dimension, indicating higher consensus. The evaluation error results are lower than those reported in the SoA [58], although it should be noted that different datasets and problems are addressed. The SoA [62] reports that in a naturalistic general content, slight disagreements in ratings can be expected, since many factors contribute to different ratings of the same stimuli, namely the subject's experience in emotion annotation, mood, personality, engagement and liking of the film, among others. The SoA reports low-to-average inter-rater agreement [14], [29], [62], increasing with the use of expert annotators [63], context [41], use of multi-modal information [41], rank annotation [6], data down-sampling [62], and outlier removal [22], [64]. After testing two alternative approaches, the authors in [36] report that for live annotation and textual annotation: *"The agreement of annotators remains very small, showing again the difficulty and inherent subjectivity of sentiment annotation"*. Our findings are in line with what is described in [62]: *"Experiments on several types of material provide information about their characteristics, particularly the ratings on which people tend to agree. Disagreement is not necessarily a problem in the technique. It may correctly show that people's impressions of emotion diverge more than commonly thought"*.

Self-report Coherence: The three segmentation algorithms lead to similar annotations when averaged across the subjects' reports for a given timestamp (see Table VII). For each timestamp, low variability is observed, below a 0.2 amplitude on a [1, 5] scale. The high coherence can be explained since the same subjects annotate similar timestamps across algorithms, resulting in the annotation of roughly the entire content, with

each timestamp being annotated by at least one subject, and in several cases by all (as seen in Table IV). The high annotation's coherence across the segmentation algorithms confirms the reliability of the retrospective annotation.

Comparison to the Reference Annotations: The standard deviation between the EmotiphAI Annotator annotations and the SoA (LIRIS-ACCEDE) (see Table VIII) showed a maximum difference of around 0.65 in a [1, 5] scale. However, the value is still below the "critical" threshold (≈ 0.71) given by [58]. No meaningful difference was detected in the valence and arousal dimensions or across the different segmentation algorithms. The SoA ground-truth was obtained in real-time and using continuous annotation $[-1, 1]$, while the EmotiphAI Annotator uses a SAM discrete scale $\{1, 5\}$ annotated retrospectively after the movie, which can introduce both latency and scale differences in the two annotations. The scale differences can be a source of error in the observed standard deviation between the two annotations. Nevertheless, overall the experimental results are in line with the SoA: *"Annotators agree on the trends but disagree on the values"* [58]; *"True emotion, however, does not necessarily fall into only one of the discrete emotion space sampling point"* [64].

Comparison to Electrodermal Activity. The EDA method outperformed the remaining, with the arousal self-reports showing lower-average to good correlation (0.26 to 0.8 correlation in [0, 1] scale) to the EDA data given by the MAP [60] in Table IX. The results reinforce the EDA as a marker of SNS activity given by the arousal dimension. The EmotiphAI Annotator Scene and EDA segmentation methods obtain competitive results comparatively with the SoA, outperforming their results for two of the videos ("Tears of Steel" and "After the Rain"). For the "Elephant's Dream" video, the lower correlation score may result from the fact that the arousal average self-report values show minimum variability with no major emotional event (self-report values around the neutral state 3 for the entire content). The low variability annotations can mean that there was no emotion elicitation or low engagement in the video. The literature [60] reports that correlations to EDA are expected when an emotional event is detected by an increase or decrease of the emotion annotations from the neutral state (3 in the EmotiphAI Annotator). Overall, the EDA method is preferred for the selection of the clips for emotion annotation, by selecting moments for emotion annotation correlated with changes in SNS activity.

C. Content Segmentation Method

Amongst the segmentation methods, the EDA-based showed advantages, namely: 1) Lower annotation effort for the user, as expressed by the lower annotation time comparatively to the remaining methods); 2) The obtained self-reports display a higher correlation with the physiological dynamics (The EDA data, as an indicator of the SNS activity, is prone to select SNS-related events for emotion annotation); 3) Scene-based segmentation is limited on the use of a movie for emotion elicitation (unlike EDA-based); and 4) Allows sorting the events by intensity, which the literature has shown to be easier to recall [52], hence being important to show first to ensure that at least these are annotated (if the subjects do not comply with completing the rating of all the segments in a real-world annotation scenario).

It should be noticed that the EDA segmentation requires additional effort (and hardware) to record EDA data, which is not necessary for the other two methods. However, when collecting physiological data for the development of emotion recognition algorithms that can be mitigated.

Overall, the annotation segmentation method should be chosen according to the needs of the research design and the goals of the study.

D. Limitations

A few limitations were identified, namely that the population used in this study was around the same age. To ensure generalisation into the real world a broader age gap should be analysed. The participant's recruitment was made through an open call within an academic population, with no intrinsic motivation.

The segmentation based on high-intensity EDA events can introduce specific challenges. Namely, it can lead to only high arousal segments being selected for labelling, potentially introducing a bias towards high-arousal moments in the dataset. Additionally, such a focus can disrupt the natural temporal sequence of the video, which may complicate the process of anchoring memories associated with these events.

The scene-based segmentation was performed using a 10-second design criterion to show the last moments of each video scene. Future work could explore the annotation of diverse scene moments.

It should be noted that the annotation platform was validated on the LIRIS-ACCEDE Continuous collection with an average movie length of 10.72 minutes. Moving forward, we are keen on exploring and validating our methodology on a novel dataset with more extended content, where further challenges can arise with longer durations.

VII. Conclusion

We introduced a novel platform for ground-truth emotion collection of naturalistic content in longer videos, the EmotiphAI Annotator. It contains built-in content segmentation algorithms that select moments of the film for annotation, reducing the required number of annotations, and simplifying the annotation process for the annotation of longer videos by more people. The use of long-duration content instead of a few minute clips (the current practice in the SoA), approximates the elicited emotional experience to the ones obtained genuinely in a stepped retrospective approach that tries to capture the best of real-time and retrospective annotation.

The platform was analysed considering its usability and annotation accuracy. The experimental results showed that the EmotiphAI Annotator provides a good user experience with a low mental workload and that the retrospective annotations can be used as a reliable ground-truth estimate for emotion recognition systems. Among the segmentation methods, the EDA demonstrated low annotation effort and showed correlation to EDA data, ensuring that SNS-related events are selected for

emotion annotation. As such, the EmotiphAI Annotator can enable quick, reliable and low mental workload emotion reports across longer elicitation content.

Future work may tackle the limitations of our proposed method, namely: 1) Acquisition of an in-the-wild database for emotion recognition using physiological data and long-duration videos; 2) Study of the relation between the EDA events and their emotion meaningfulness; and 3) Exploration of diverse scene moments for annotation. We foresee that EmotiphAI Annotator could be the basis for widespread emotion annotation across diverse distributed real-world scenarios, leading to the creation of large databases for emotion recognition.

REFERENCES

[1] M. Adheena, N. Sindhu, and S. Jerritta, "Physiological detection of anxiety," in *Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol.*, 2018, pp. 1–5.

[2] L. B. Krithika and P. G. G. Lakshmi, "Student emotion recognition system for e-learning improvement based on learner concentration metric," *Procedia Comput. Sci.*, vol. 85, pp. 767–776, 2016.

[3] A. Alrihaili, A. Alsaedi, K. Albalawi, and L. Syed, "Music recommender system for users based on emotion detection through facial features," in *Proc. Int. Conf. Developments eSystems Eng.*, 2019, pp. 1014–1019.

[4] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019.

[5] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *Proc. ISCA Tutorial Res. Workshop Speech Emotion*, 2000, pp. 19–24.

[6] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 574–580.

[7] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, *RCEA: Real-Time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels*. New York, NY, USA: ACM, 2020, pp. 1–15.

[8] K. Sharma, C. Castellini, F. Stulp, and E. L. van den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 78–84, First Quarter 2020.

[9] F. Larradet, R. Niewiadomski, G. Barresi, D. Caldwell, and L. S. Mattos, "Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection," *Front. Psychol.*, vol. 11, 2020, Art. no. 1111.

[10] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proc. Int. Conf. Multimodal Interaction*, New York, NY, USA: ACM, 2020, pp. 784–789.

[11] W. Boucsein, *Principles of Electrodermal Phenomena*. Boston, MA, USA: Springer, 2012, pp. 1–86.

[12] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Second Quarter 2021.

[13] A. Zlatintsi et al., "COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP J. Image Video Process.*, vol. 271, 2017, Art. no. 54.

[14] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 77–83.

[15] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionML," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 709–710.

[16] P. Bota, C. Wang, A. Fred, and H. Silva, "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?," *Sensors*, vol. 20, no. 17, 2020, Art. no. 4723.

[17] S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, First Quarter 2012.

[18] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, First Quarter 2012.

[19] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Second Quarter 2018.

[20] J. A. Healey and R. W. Picard, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, Massachusetts Ins. Technol., Cambridge, MA, USA, 2000.

[21] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves, "The emotional movie database (EMDB): A self-report and psychophysiological study," *Appl. Psychophysiol. Biofeedback*, vol. 37, no. 4, pp. 279–294, 2012.

[22] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Third Quarter 2015.

[23] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. Int. Conf. Multimodal Interaction*, NY, USA:ACM, 2018, pp. 400–408.

[24] C. Y. Park et al., "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 293.

[25] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A database for cognitive load, affect and stress recognition," in *Proc. Int. Conf. Biomed. Innovations Appl.*, 2019, pp. 1–4.

[26] Y. Liu, T. Gedeon, S. Caldwell, S. Lin, and Z. Jin, "Emotion recognition through observer's physiological signals," 2020, *arXiv:2002.08034*.

[27] P. Lopes, G. N. Yannakakis, and A. Liapis, "RankTrace: Relative and unbounded affect annotation," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 158–163.

[28] J. M. Girard and A. G. C. Wright, "DARMA: Software for dual axis rating and media annotation," *Behav. Res. Methods*, vol. 50, no. 3, pp. 902–909, 2018.

[29] D. Melhart, A. Liapis, and G. N. Yannakakis, "PAGAN: Video affect annotation made easy," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 130–136.

[30] J. Broekens and W.-P. Brinkman, "AffectButton: A method for reliable and valid affective self-report," *Int. J. Hum.- Comput. Stud.*, vol. 71, no. 6, pp. 641–667, 2013.

[31] T. Baur et al., "eXplainable cooperative machine learning with NOVA," *KI-Künstliche Intelligenz*, vol. 34, pp. 143–164, 2020.

[32] N. Runge, M. Hellmeier, D. Wenig, and R. Malaka, "Tag your emotions: A novel mobile user interface for annotating images with emotions," in *Proc. Int. Conf. Human-Comput. Interaction Mobile Devices Serv. Adjunct*, New York, NY, USA: ACM, 2016, pp. 846–853.

[33] G. F. D. Salvador, P. J. Bota, V. Vinayagamoorthy, H. Plácido da Silva, and A. Fred, *Smartphone-Based Content Annotation for Ground Truth Collection in Affective Computing*. New York, NY, USA: ACM, 2021, pp. 199–204.

[34] E. Gatti, E. Calzolari, E. Maggioni, and M. Obrist, "Emotional ratings and skin conductance response to visual, auditory and haptic stimuli," *Sci. Data*, vol. 5, no. 1, 2018, Art. no. 180120.

[35] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[36] T. Schmidt, I. Engl, D. Halbhuber, and C. Wolff, "Comparing live sentiment annotation of movies via arduino and a slider with textual annotation of subtitles," in *Proc. Conf. Digit. Humanities Nordic Countries*, 2021, pp. 212–223.

[37] M. Robinson and G. Clore, "Belief and feeling: Evidence for an accessibility model of emotional self-report," *Psychol. Bull.*, vol. 128, pp. 934–60, 2002.

[38] L. F. Barrett, "The relationships among momentary emotion experiences, personality descriptions, and retrospective ratings of emotion," *Pers. Social Psychol. Bull.*, vol. 23, no. 10, pp. 1100–1110, 1997.

[39] C. Röcke, C. A. Hoppmann, and P. L. Klumb, "Correspondence between retrospective and momentary ratings of positive and negative affect in old age: Findings from a one-year measurement burst design," *J. Gerontol.: Ser. B*, vol. 66B, no. 4, pp. 411–415, 2011.

[40] E. Öhman, "Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task," in *Proc. Digit. Humanities Nordic Countries Conf.*, 2020, pp. 293–301.
[41] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human-computer interaction: Comparison and methodological improvements," *J. Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.
[42] A. Mill, A. Realo, and J. Allik, "Retrospective ratings of emotions: The effects of age, daily tiredness, and personality," *Front. Psychol.*, vol. 6, 2016, Art. no. 2020.
[43] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.
[44] H. L. Urry and J. J. Gross, "Emotion regulation in older age," *Curr. Directions Psychol. Sci.*, vol. 19, no. 6, pp. 352–357, 2010.
[45] E. Schryer and M. Ross, "Evaluating the valence of remembered events: The importance of age and self-relevance," *Psychol. Aging*, vol. 27, pp. 237–42, 2011.
[46] D. L. Paulhus and O. P. John, "Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives," *J. Pers.*, vol. 66, no. 6, pp. 1025–1060, 1998.
[47] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cogn. Emotion*, vol. 23, no. 2, pp. 209–237, 2009.
[48] R. D. Lane, G. L. Ahern, G. E. Schwartz, and A. W. Kaszniak, "Is alexithymia the emotional equivalent of blindsight?," *Biol. Psychiatry*, vol. 42, no. 9, pp. 834–844, 1997.
[49] P. Bota, E. Flety, H. P. D. Silva, and A. Fred, "Emotiphai: A biocybernetic engine for real-time biosignals acquisition in a collective setting," *Neural Comput. Appl.*, vol. 35, no. 8, pp. 5721–5736, 2023.
[50] W. Boucsein et al., "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012.
[51] W. Boucsein, *Methods of Electrodermal Recording*. Boston, MA, USA: Springer, 2012, pp. 87–258.
[52] M. Mather, "Emotional memory," in *The Encyclopedia of Adulthood and Aging*, S. K. Whitbourne, Ed., USA: John Wiley & Sons, Inc., 1st ed., 2016, doi: 10.1002/9781118528921.wbeaa243.
[53] C. Liu, A. Shmilovici, and M. Last, "Towards story-based classification of movie scenes," *PLoS One*, vol. 15, no. 2, pp. 1–22, 2020.
[54] D. Batista, H. Plácido da Silva, A. Fred, C. Moreira, M. Reis, and H. A. Ferreira, "Benchmarking of the BITalino biomedical toolkit against an established gold standard," *Healthcare Technol. Lett.*, vol. 6, no. 2, pp. 32–36, 2019.
[55] J. Brooke, "SUS: A retrospective," *J. Usability Stud. Arch.*, vol. 8, pp. 29–40, 2013.
[56] R. A. Grier, "How high is high? A meta-analysis of NASA TLX global workload scores," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 59, no. 1, pp. 1727–1731, 2015.
[57] J. R. Lewis and J. Sauro, "Item benchmarks for the system usability scale," *J. Usability Stud.*, vol. 13, no. 3, pp. 158–167, 2018.
[58] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2005, pp. 381–385.
[59] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2376–2379.
[60] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Continuous arousal self-assessments validation using real-time physiological responses," in *Proc. Int. Workshop Affect Sentiment Multimedia*, New York, NY, USA: ACM, 2015, pp. 39–44.
[61] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 73–78.
[62] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *Int. J. Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, 2012.
[63] C. C. Musat, A. Ghasemi, and B. Faltings, "Sentiment analysis using a novel human computation game," in *Proc. Workshop People's Web Meets NLP: Collaboratively Constructed Semantic Resour. Their Appl. NLP*, 2012, pp. 1–9.
[64] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. IEEE Proc. Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

Patrícia Bota (Student Member, IEEE) received the MSc degree in biomedical engineering from the NOVA University of Lisbon. She is currently working toward the PhD degree with Instituto Superior Técnico, where she is involved with the study of group human emotions in the wild through artificial intelligence algorithms based on physiological data. Her main research interests include artificial intelligence, affective computing, physiological data, and signal processing.

Pablo Cesar (Senior Member, IEEE) leads the Distributed and Interactive Systems Group, Centrum Wiskunde & Infomartica (CWI) and is a professor with TU Delft, The Netherlands. His research combines human-computer interaction and multimedia systems and focuses on modelling and controlling complex collections of media objects (including real-time media and sensor data) that are distributed in time and space. He has recently received the prestigious 2020 Netherlands Prize for ICT Research because of his work on human-centred multimedia systems. He is the Principal investigator from CWI on several projects on social virtual reality and affective computing. He is a member of the Editorial Board of the *IEEE Multimedia*, *ACM Transactions on Multimedia,* and *IEEE Transactions of Multimedia*, among others. He has acted as an Invited Expert at the European Commission's Future Media Internet Architecture Think Tank.

Ana Fred (Member, IEEE) received the MS and PhD degrees in electrical and computer engineering (ECE) from IST, in 1989 and 1994, respectively. She has been a faculty member with IST since 1986, where she has been a professor with the Department of ECE, and more recently with the Department of Biomedical Engineering. She is a senior researcher with IT. Her main research interests include signal processing, pattern recognition and machine learning. She has done pioneering work on clustering, namely on cluster ensemble approaches. She has published more than 200 papers in international refereed conferences, peer-reviewed journals, and book chapters.

Hugo Plácido da Silva (Senior Member, IEEE) is currently working toward the PhD degreee in electrical and computer engineering from the IST-UL. He is an award-winning inventor, researcher, and entrepreneur, having co-founded multiple technology-based companies in the field of biomedical engineering. He has been a Researcher at IT since 2004 and a faculty member of IST-UL since 2019. His current interests include biosignal research, system engineering, signal processing, and machine learning, in which he holds 7 patents and has been distinguished with numerous academic and technical awards. The most recent distinction was the prestigious IEEE Entrepreneurship Impact Award 2023, which recognizes an individual who has had a significant impact on the engineering-driven entrepreneurial ecosystem.