Improving Continuous Emotion Annotation in Video Platforms via Physiological Response Profiling

Swarnali Banik*, Sougata Sen*, Snehanshu Saha*, Surjya Ghosh*
*Department of Computer Science and Information Systems, BITS Pilani Goa, INDIA
Email: {p20210016, sougatas, snehanshus, surjyag}@goa.bits-pilani.ac.in

***Abstract*—Many video applications (e.g., gaming, meeting, tutoring) aim to improve the user's interaction experience based on continuously inferred user emotion. To infer user emotion, these apps typically deploy machine learning models, trained with continuously collected emotion ground truth labels. However, as continuous annotations are generally collected as emotion self-reports during video consumption (using an auxiliary device), they incur significant annotation effort. To address this problem, we propose PResUP, a framework that creates users' profile using physiological responses (e.g., GSR or galvanic skin response) and deploys an LSTM network to identify the *opportune* probing moments for emotion ground truth (emotion self-report) collection instead of continuous annotation. We evaluate the proposed approach on a large-scale publicly available dataset (CASE) containing the physiological signals of subjects during video consumption. The evaluation of PResUP reveals that it reduces the probing rate by 30.3% (on average), detects the opportune probing moments with a TPR (True Positive Rate) of 86.1%, and yet maintains the quality of the emotion annotations as observed in the continuous self-reports. Furthermore, we evaluated the generalizability of PResUP on another public dataset (K-emocon); which reveals an average probing rate reduction of 25.71%. These results underscore the efficiency of PResUP in reducing the continuous emotion annotation overhead.**

***Index Terms*—Continuous emotion annotation, Video application, Physiological response profile**

I. Introduction

Many video-based applications (e.g., online meeting platforms [1], online tutoring apps [2], entertainment platform [3]) aim to create a personalized and interactive user experience based on user emotion (e.g., stress, anxiety). These applications generally use machine learning models that infer emotions continuously based on the variations in the physiological responses during user interaction. To train such fine-grain emotion inference models, the emotion ground truth labels must also be collected continuously [4]. But, as emotion ground truth labels are typically collected as self-reports, the continuous emotion annotation process during video consumption becomes challenging because the users need to focus on two tasks at the same time – (i) watching the video, and (ii) providing the emotion self-report continuously. As a result, the user's viewing experience degrades, and the cognitive load increases. Therefore, efficient approaches to reduce the continuous emotion annotation effort are essential.

In the existing literature, researchers use auxiliary devices like mice [5] or joysticks [6] for continuous emotion annotation during video watching. Examples include the CASE dataset [7] and FEELTRACE [5]. However, these approaches have same underlying challenges – concentrating on video consumption and annotating emotion at the same time. More recently, Park et al. proposed a post-interaction emotion annotation approach at every 5-second interval [8]. However, this is burdensome for long videos and suffers from recall bias.

The aforementioned challenges of continuous annotation can be overcome considering several factors. First, as human emotion persists for some time (defind as persistence time [9]) once experienced and the emotional content does not change very frequently (e.g., every consecutive frame) in a video, it may not be required to collect emotion annotations continuously. Second, as the physiological signals vary under emotional influence [10], [11], it may be worth capturing the emotion self-reports only at these *opportune* moments, when physiological signals vary substantially. Finally, these opportune moments of physiological signal variations can be captured using the state-of-the-art sequential networks (e.g., LSTM - Long short-term memory) based on the time-series properties of these signals [12], [13].

We, in this paper, propose the PResUP (**P**hysiological **Res**ponse based **U**ser **P**rofiling) framework for opportunistic continuous emotion annotation based on the physiological response profile of users leveraging the aforementioned intuitions. The framework operates in the following phases (Section V). First, it creates the user profile, quantifying the variations in the physiological responses during different segments of the video. Intuitively, higher the variation in the physiological responses, more opportune the segment is for emotion self-report collection (as emotional stimuli influences the physiological responses [10]). The physiological response profile contains the summary statistics of user's physiological response behavior during the opportune and inopportune probing moments. This quantification helps in the second phase, when a group of similar users are identified using k-means algorithm [14] based on the profile similarity. In the last phase, we enable data sharing among similar users to train an LSTM-based sequential network to detect the opportune moments for emotion self-report collection. The process of clustering similar users and sharing data among them helps to train the LSTM model with a larger pool of data for better performance.

We evaluate PResUP on the publicly available CASE dataset [7], which captures the physiological responses from six sensor streams from 30 participants as they watch eight different videos and perform the continuous annotation using a joystick (Section III). We segment the sensor data into small

2156-8111/24/$31.00 ©2024 IEEE
DOI 10.1109/ACII63134.2024.00047

windows, compute the physiological response variations (in these segments) using the RuLSIF algorithm [15], and tag a segment as opportune (or not) based on the degree of variation in physiological response. We find a causal relationship (Section IV) between emotion variation and physiological signal change, which reinforces that the opportune segments are ideal for probing the user for emotion self-reports. The empirical evaluation on this dataset reveals that PResUP reduces the average probing rate by 30.3% (with respect to baselines), detects the opportune probing moments with high accuracy (True Positive Rate: 86.1%, and False Positive Rate: 12.2%), and yet maintains the emotion annotation quality as observed in continuous annotations. A follow-up evaluation of PResUP on another public dataset (K-emocon) reveals an average probing rate reduction of 25.71%, thus underscoring its generalizability across datasets. In summary, the major contributions of this paper are as follows,

- We propose a framework PResUP (Section V) that exploits the physiological response behavior to reduce the continuous emotion annotation effort. It encompasses an LSTM network to identify the opportune probing moments for emotion self-report collection instead of continuous emotion annotation.
- We propose a physiological response profile construction approach (Section V-A) that allows to quantify user behavior based on physiological response. The response profile allows to identify similar users (Section V-B) so that LSTM network embedded in the PResUP framework can be trained more efficiently (Section V-C).
- Finally, we evaluate PResUP on publicly available CASE dataset (Section VI) and conduct a generalizability study on another dataset (K-emocon, Section VII) to demonstrate its efficiency in reducing the probing rate (i.e., annotation effort) without compromising the annotation quality.

II. Related Works

In the existing literature, emotion annotations are typically collected through self-reports using survey questionnaires like the Self-assessment Manikin (SAM) after watching a video [16]. This method fails to capture intra-video emotion variations. To address this problem, researchers use continuous annotation strategies where participants provide emotion annotations using a mouse [5] or joystick [6]), as they watch the video. For example, in the CASE (*Continuously Annotated Signals of Emotion*) dataset [7], participants continuously provided emotion annotation (*valence* and *arousal* based on the Circumplex Model of emotion [17]) using a joystick. Works such as FEELTRACE [5], GTrace [18], and DARMA [19] have also collected continuous emotion annotations using similar devices. But these approaches require users to annotate during video consumption. To overcome this, Park et al. proposed a post-interaction emotion annotation approach during which the annotations are collected at every 5-second based on audio-visual recordings [8]. However, in this approach, the annotation cost would be high for long videos and there may be an impact of recall bias due to post-interaction nature [20].

The rich literature in affective computing suggests a strong relationship between human emotions and physiological signals [10]. Physiolgical signals (e.g., Galvanic skin response (GSR), Electromyography (EMG), and Respiratory signal) vary once an emotional stimuli (such as emotion inducing video, audio) is applied [21], [22]. Therefore, the substantial variation in physiological responses can be an indicator of emotional change.

In summary, fine-grain continuous emotion annotation is challenging. Rather, probing a user for annotation at opportune moments based on the variation in the physiological responses can be a possible option. However, such an approach has not been investigated in the case of continuous emotion annotation, which we investigate in this paper.

III. CASE Dataset

The CASE (Continuously Annotated Signals of Emotion) dataset is a publicly available dataset [7]. It captures continuous emotion annotations and physiological responses of 30 volunteers (15F, 15M) aged between 22 and 37 years as they watch eight different stimuli videos (as present in the dataset) that embed different emotions from the Circumplex model (Fig. 1). For example, video 1 and 2 evoke an amusing emotion (having high valence and arousal); video 3 and 4 induce boredom (having low valence and arousal); video 5 and 6 embed relaxed emotion (having high valence and low arousal); and video 7 and 8 elicit scary emotion (having low valence and high arousal). Notably, the eight stimuli videos were selected in such a manner so that there are two videos from each quadrant (therefore, the entire spectrum of valence and arousal were covered during the data collection).

The emotions were recorded continuously (during video consumption) using a joystick based on the Circumplex model of emotion [17]. As per this model (see Fig. 1), human emotion comprises of valence and arousal, accordingly, the annotations were collected using the Joystick on a 2D plane to record valence and arousal scores (on a scale of 1 to 9). In parallel, following physiological signals were continuously collected – Electrocardiograph (ECG), Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Respiration (RSP), Skin Temperature (SKT), and Electromyography (EMG).

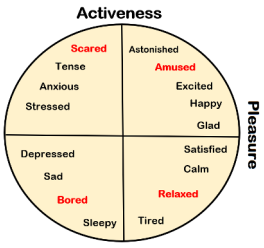


Fig. 1: Circumplex model of emotion [17]. The stimuli videos embed following emotions (amused, scared, bored, relaxed) from different quadrants to cover all combinations of valence and arousal.

A. Data Preprocessing

We performed the following pre-processing steps,

1) Segmenting Physiological Responses: First, we segment the physiological responses into fixed size windows. Although the dataset contains physiological responses from different signals (BVP, ECG, EMG, GSR, RSP, and SKT), we discarded the ECG signal as it is relatively noisy [23]. For every participant and video combination, we segment the remaining physiological signals into 5-seconds windows (this window length also adopted in earlier works [8], [24]).

2) Labeling Opportune Moments: Next, we label the segments as opportune (or not) leveraging the physiological signal changes as done in earlier work [24]. Broadly, the labeling algorithm operates in two steps. In the first step, it computes change point score for every segment using the RuLSIF algorithm. The algorithm computes the divergence between consecutive signal segments as the change point score [25]. A high degree of divergence is denoted by a high change point score and vice versa. In the second step, segments with high change point scores are clustered and marked as opportune.

3) Dataset Description: After the pre-processing steps, we obtain ≈10.125 hours of physiological response and emotion annotation data (from all users) that spans across a total of 7290 segments (5 seconds each). Among these segments, 11.6% are opportune, and 88.4% are inopportune. Each user contributes to 243 segments (as every user watches the same videos). On average, there are 28.3 (SD: 4.71) opportune segments and 214.7 (SD: 4.71) inopportune segments per user.

IV. Feasibility Study: Causality Analysis

We performed a feasibility study to test the hypothesis that emotion variations cause physiological signal variations. If this hypothesis is found to be true, then by probing at the opportune moments (denoted by substantial change in the physiological signals), the emotion variations can be captured.

Causality between Emotion and Physiological Response: Emotion variations are reflected in valence (or arousal) scores, whereas the physiological signal variations are manifested in the change point scores. Therefore, we investigate the causal relationship between valence-arousal scores and change point scores. In specific, we perform Granger causality test [26] to test if variation in valence and arousal causes variation in change point score. This test is a statistical hypothesis test for determining whether one time series (in this case emotion scores) is useful in forecasting another (in this case change point score). Mathematically it is expressed as follows. One time series X_t does "Granger-causes" another time series Y_t, if the past values of X_t help to predict the future values of Y_{t+1}. Granger causality or G-causality works on the principle of modeling the two processes X and Y as auto-regressive processes. Specifically, in order to determine if '*Y G-causes X*', the two models considered are,

$$X(t) = \sum_{\tau=1}^{\infty}(p_\tau X(t-\tau)) + \sum_{\tau=1}^{\infty}(r_\tau Y(t-\tau)) + \varepsilon_c, \quad (1)$$

$$X(t) = \sum_{\tau=1}^{\infty}(q_\tau X(t-\tau)) + \varepsilon, \quad (2)$$

where t stands for time, p_τ, q_τ, r_τ are coefficients at a time lag of τ and $\varepsilon_c, \varepsilon$ are error terms. Covariance stationarity is assumed for both X and Y. Whether Y G-causes X (or not) can be predicted by the measure known as F-statistic which is the log ratio of the prediction error variances:

$$F_{Y\to X} = \ln \frac{var(\varepsilon)}{var(\varepsilon_c)}. \quad (3)$$

If the model represented by equation (1) is a better model for $X(t)$ than equation (2), then var(ε_c) < var(ε) and $F_{Y\to X} > 0$, suggesting that Y *Granger causes* X. Even though G-causality uses the notion of autoregressive models for the variables, the generic nature of this modeling with minimal assumptions about the underlying mechanisms makes it a very popular choice in a wide range of disciplines [27], [28].

Causality Test Outcome: We perform this test twice (once for arousal and once for valence). We accumulate the arousal scores and the change point scores for every segment in two different lists and run the causality test [29]. The test outputs the following results ($F = 8.1140$, $\chi^2 = 24.3654$, p-val = 0.0000, df = 3). The null hypothesis is that arousal scores do not Granger cause the change point scores. However, as per the results ($p < 0.05$), we reject this null hypothesis thus implying a causal relationship between arousal and change point scores. We repeat the same steps for valence, but in this case we are unable to reject the null hypothesis ($F = 0.7022$, $\chi^2 = 2.1087$, p-val = 0.5502, df = 3). This can be attributed to the fact that as arousal relates with activeness (whereas valence corresponds to pleasantness), variations in arousal cause significant changes in the physiological signals (unlike valence). Nevertheless, this finding highlights that emotional variations reflect in physiological signal changes (as emotion comprises of both valence and arousal). Therefore, by probing at the opportune segments, we can capture the emotional changes. So, we aim to detect the opportune moments (using PResUP framework) as discussed next.

V. PResUP Framework

In this section, we present the overview of the PResUP framework (Fig. 2). It operates in three phases - (a) physiological response profile creation, (b) similar user identification, (c) opportune moment detection. First, we create user profile based on the physiological responses during opportune and inopportune segments. Next, similar users are clustered based on the physiological response profile. Finally, cluster-specific LSTM models are constructed for opportune moment detection. We discuss each of the phases in detail now.

A. Physiological Response Profile Creation

The physiological response profile creation for a user is a two-step process. In the first step, we compute the change point scores between every two consecutive segments (s_i, s_{i+1}, where $1 \le i \le n-1$; n is the total number of segments). In the second step, we compute the mean (μ) and standard deviation (σ) of the change point scores for both opportune and inopportune segments. The user profile is constructed as a

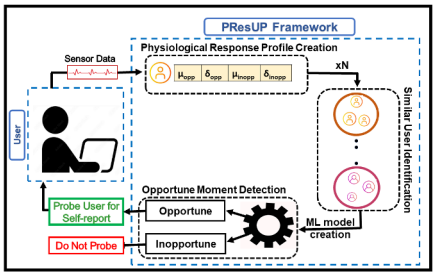
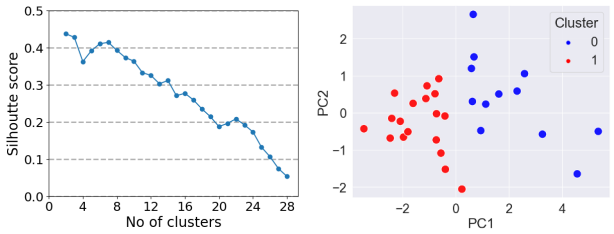


Fig. 2: Overview of the PResUP framework's three phases - (a) physiological response profile creation (b) similar user identification (c) opportune moment detection. In the first phase (during video consumption), based on the physiological responses, the user response profile is created. In the second phase, the response profiles are used to cluster similar users. In the final phase, cluster-specific LSTM models are created for opportune probing moment detection.

four tuple $<\mu_{opp},\ \sigma_{opp},\ \mu_{inopp},\ \sigma_{inopp}>$, where $\mu_{opp},\ \sigma_{opp}$ indicate the mean and std. dev at the opportune segments, and $\mu_{inopp},\ \sigma_{inopp}$ indicate the mean and std. dev at the inopportune segments. These steps are repeated for every user. Since the profile captures two types of change point scores, it allows to identify users having similar physiological response variations during opportune and inopportune moments.

B. Similar User Identification

This phase identifies similar users based on the physiological response profile. We run k-means algorithm on the profile vectors for different values of k. We vary the value of k from 2 to 28, and note the cluster quality in Fig. 3a using Silhoutte score [30]. We obtain the highest value of Silhoutte score for k = 2, therefore, we group the users into two clusters. To visualize the user groups, we perform PCA (Principal Component Analysis) [31] on the profile vectors and display the clusters on a 2D plane (Fig. 3b). We observe that there are almost equal number of users in both the clusters. Once we cluster the users, we construct the opportune moment detection model by sharing data among the similar users.



(a) Silhoutte score variation **(b)** User clustering (k=2)

Fig. 3: Clustering users based on the physiological response profile vectors - (a) variation in Silhoutte score for different number of clusters (k), which reveals that number of optimal clusters is two (b) visualization of group of similar users for optimal number of cluster (k=2) after applying PCA on the profile vectors

C. Opportune Moment Detection

To detect the opportune probing moments, we develop LSTM-based models by sharing data among users present in a cluster. We show the model architecture in Fig. 4. Given a sequence of segments $\mathcal{X} = \langle \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_T \rangle$ containing physiological responses, the LSTM model takes an input $\mathbf{x}_t = [chp_t,\ gsr_t,\ val_{video},\ aro_{video}]^T$ at each step t, where chp_t is the change point score at time t, gsr_t is the difference in the mean value of GSR between segment t and $t-1$; val_{video} and aro_{video} denote the valence and arousal of the video respectively. These two values ($val_{video},\ aro_{video}$) remain same in every segment for a given stimuli video depending on the valence and arousal of the video. For example, for stimuli video 1, the valence and arousal values are '1' as this video embeds the emotion amusing (as noted in Section III), which belongs to the first quadrant of the Circumplex model (Fig. 1). The other features ($chp_t,\ gsr_t$) capture the temporal changes in physiological signals with emotion variation. We noted that using GSR as a feature, the opportune moment detection performance improves (similar to earlier works on inferring emotional engagement [32]). So, we use this as a model input. However, as other physiological signal (BVP, RSP, SKT, EMG) do not improve the opportune moment detection performance, we do not use them.

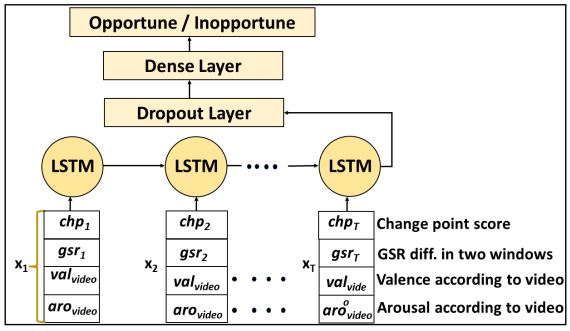


Fig. 4: LSTM-based architecture used in the PResUP framework

The proposed architecture consists of T LSTM cells (we used T=10, as we found empirically it returns superior performance).

The LSTM embedding is input to a dropout layer, followed by a dense (fully connected) layer. The output is finally connected to a sigmoid function to perform a binary classification i.e., opportune or inopportune. Therefore, we use binary crossentropy loss for training this architecture.

We specify the hyperparameter details in Section VI-A3. While training the model for a particular user u, we use data from *only* those users, who belong to the same cluster as u (Fig. 3b, Section V-B). We do not use any data from user u for training the model, rather it is used for evaluating the model. The same process is repeated for all the users.

VI. Evaluation

We evaluate the performance of PResUP in reducing the annotation effort (measured in terms of probing rate), detecting the opportune moments, and maintaining the annotation quality. Accordingly, we compare the performance of the PResUP with a set of baselines using the metrics defined below.

A. Experiment Setup

The experiment setup for evaluation of PResUP is as follows. First, we construct the physiological response profile of the users and cluster them. Next, for every user (u), we train the proposed LSTM model using data of other users present in that cluster only and test using the left-out user's (u) data. The process is repeated for each user. The number of segments detected as opportune by the model is considered as the number of probes answered by a user.

1) Baseline: We discuss the baselines below used for comparing the performance of the PResUP framework.

- **Time-based Probing Strategy (TBS) [8]:** This baseline collected the annotation at a fixed interval (every 5 sec., based on earlier work [8]).We compare PResUP with this baseline as it demonstrates the efficiency of the opportunistic probing over continuous (fixed interval) probing.
- **Feature-based Probing Strategy (FBS):** In this baseline, we train a classical machine learning model (not LSTM) by adopting the same approach (user profiling, followed by clustering) as used in PResUP. However, We concentrate on one window at a time (unlike PResUP), extracting the same features. Notably, Multiple algorithms (RF, SVM, XGBoost) were implemented, with XGBoost selected for its superior performance. This comparison underscores the utility of temporal sequence as considered in PResUP.
- **Personalized Probing Strategy (PPS):** This baseline implements an LSTM model like PResUP. However, the model is *personalized* (i.e., user-specific). It is trained on 80% on the *initial* data and tested on the remaining 20% data for each user. This comparison highlights generalizability of PResUP (i.e, performance in absence of personal training data).
- **Generalized Probing Strategy (GPS):** In this case, we train an LSTM network as used in PResUP. It does not employ user profile creation and clustering (like PResUP) and is evaluated via leave-one-subject-out cross-validation. This baseline emphasizes the advantage of profile-based clustering over aggregating data from all users.
- **Age-based Probing Strategy (APS):** In this baseline, participants were clustered by age (20-24, 25-29, 30-34, 35-39 years; as mentioned in the dataset). Then cluster-specific models were implemented using XG Boost. We evaluate one user at a time and train the model with same-cluster users' data. This comparison shows the superiority of profile-based clustering over age-based one.
- **Gender-based Probing Strategy (GBPS):** This is similar to the **APS** baseline, but clustering is done based on gender (male, female). This comparison emphasizes the efficiency of profile-based clustering over gender-based clustering.
- **RNN-based Probing Strategy (RNNPS):** This baseline implements a recurrent neural network layer followed by a dropout layer (rate: 0.5) employing the same method as PResUP (user profiling and clustering). Two dense layers with ReLU activation are added, with the final layer facilitating binary classification using sigmoid activation.
- **GRU-based Probing Strategy (GRUPS):** This baseline employed a gated recurrent unit (GRU) architecture, followed by two dense layers with ReLU activation. This method adopts the same approach (user profiling and clustering) as used in PResUP. A single neuron output layer facilitated binary classification.
- **1D-CNN based Probing Strategy (CNNPS):** This baseline employs a 1D-CNN with a convolutional layer (8 filters, kernel size 3) followed by max-pooling (pool size: 2). This is followed by flattening and two fully connected dense layers with ReLU activation, finally ending with a sigmoid output layer for binary classification. This method follows the same user profiling and clustering approach utilized in PResUP.

The hyperparameter details of the baselines (**RNNPS**, **GRUPS**, **CNNPS**) are outlined in Section VI-A3.

2) Performance Metrics: We use the following metrics to evaluate the performance of the PResUP framework.

- **Probing Rate:** We compute the average number of probes issued per video for every user as the probing rate.
- **Opportune Moment Detection Metrics:** We use the standard metrics to measure the probing moment detection performance. **True Positives (TP):** Opportune moments that are correctly identified. **False Positives (FP):** Inopportune moments that are identified as opportune. **True Negatives (TN):** Inopportune moments that are identified as Inopportune. **False Negatives (FN):** Opportune moments that are identified inopportune. We calculate: **True Positive Rate (Recall or Sensitivity) TPR(%)** = $\frac{TP}{TP+FN}$ x 100. **True Negative Rate (Specificity) TNR(%)** = $\frac{TN}{TN+FP}$ × 100. **False Positive Rate (Fall-out) FPR(%)** = $100 - TNR(\%)$. **Likelihood Ratio for opportune moments (LR+)** = $\frac{TPR(Sensitivity)}{FPR(1-Specificity)}$.
- **Annotation Quality:** We investigate if there is any statistically significant difference (significance level, $\alpha = 0.05$) in the valence (and arousal) between the ground truth annotations (continuous) and the annotations collected by probing at the opportune moments.

3) Hyperparameters: To select the optimal values of the hyperparameters (for the LSTM model of the PResUP), we performed grid search. We tried with the following (i) batch sizes $(8, 16, 24)$, (ii) epochs $(25, 35, 45, 55)$ (iii) dropout rates $(0.2, 0.3, 0.4, 0.5, 0.6)$ (iv) number of LSTM nodes $(10, 20, 50, 100, 200)$ and (v) number of dense layer nodes $(10, 20, 50, 100, 200)$. We observed that for the batch size of 16, the epoch of 35, dropout of 0.5, number of LSTM nodes of 10, and number of dense nodes of 100, the best classification accuracy is obtained. For the baseline models (**RNNPS**, **GRUPS**, **CNNPS**), optimal performance was also achieved with a batch size of 16, 35 epochs, and 200 dense nodes, which were then fixed as the baseline hyperparameters.

B. Probing Rate Reduction

We present the comparison of probing rate between PResUP and baselines in Table I. PResUP has the best probing rate (on average 5.80 probes per video per user), while the **(TBS)** has the worst (30.38). The **TBS** performs the worst as it

continuously probes every 5 seconds. **PPS** also has a high probing rate (20.38) because it relies only on personal self-reports to create the opportune moment detection model. The **(FBS)** and **(GPS)** have comparatively lower probing rates of 6.50 and 6.37, respectively. As these baselines share data from other users, they reduce the individual user's probes but could not outperform the PResUP. Although the demographic-based baselines have a relatively lower probing rate (**APS**: 6.10, **GBPS**: 6.07), they still have a higher probing rate than that of PResUP. The recurrent neural network-based baselines (**RNNPS**: 6.01, **GRUPS**: 6.03) return comparatively lower probing rates than the traditional ML model but not lower than the PResUP. The **CNNPS** has a probing rate of 6.52, still inferior to PResUP.

In summary, we note that PResUP reduces the probing rate in comparison to all the baselines (maximum reduction of 80.9% wrt **TBS**, minimum reduction of 4.4% wrt **GBPS**, and an average reduction of 30.3%); but whether this reduction in probing rate influences probing moment detection performance (or the emotion annotation quality) is investigated next.

	Probing rate↓	TPR (%) ↑	FPR (%) ↓	LR+ ↑
TBS	30.38 (0.00)	100.00 (0.00)	100.00 (0.00)	1.00
PPS	20.38 (1.33)	69.21 (32.34)	16.16 (18.72)	4.28
FBS	6.50 (3.02)	70.49 (22.49)	15.02 (8.31)	4.69
GPS	6.37 (4.10)	81.30 (21.24)	14.93 (14.15)	5.45
APS	6.10 (4.04)	71.18 (20.86)	14.35 (15.98)	4.96
GBPS	6.07 (4.07)	82.12 (16.18)	13.68 (13.19)	6.00
RNNPS	6.01 (8.72)	82.66 (11.52)	12.88(5.24)	6.41
GRUPS	6.03 (8.98)	76.14 (10.75)	13.68 (5.37)	5.56
CNNPS	6.52 (6.76)	82.55(10.39)	15.47(6.96)	5.33
PResUP	**5.80 (2.89)**	**86.07 (11.64)**	**12.27 (9.67)**	**7.01**

TABLE I: Performance comparison of PResUP and baselines. The values indicate the user-wise average and std.dev (inside parenthesis) for a metric. PResUP outperforms all baselines in terms of probing rate (least probing rate), LR+ (highest LR+), TPR, and FPR. Although **TBS** has the highest TPR, it also has the worst FPR (100%). ↑, ↓ indicate higher and lower value preferred respectively.

C. Opportune Probing Moment Detection

We next evaluate the probing moment detection performance of PResUP. First, we compare the model performance using the likelihood ratio (LR+) in Table I. Intuitively, LR+ ($= \frac{TPR}{FPR}$) should be high enough to instill confidence in detecting the opportune moments. Notably, a very high value of LR+ implies that if the test is positive, the condition of opportune moment is definitely present. Therefore, the model with higher LR+ is desired. We observe that PResUP outperforms (LR+ = 7.01) all the baselines and **TBS** performs the worst (LR+ = 1). Next, we delve deep to investigate if PResUP also has a high TPR and low FPR.

1) TPR Analysis: We present the comparison of mean TPR in detecting the opportune probing moments in Table I. The **PPS** performs the worst (mean TPR: 69.21%) followed by the **(FBS)** (mean TPR: 70.49%). This highlights that relying only on the personal data to train the opportune moment detection model (as done in the **PPS**) does not yield good performance. Similarly, developing a feature-based model **(FBS)** without considering the temporal sequence is also not a good option. However, the sequence-based model **(GPS)** that aggregates data from all users (excluding the physiological response similarity) returns comparatively better performance (mean TPR: 81.30%). The demographic-based baselines (**APS, GBPS**) return relatively better performance with mean TPR of 71.18% and 82.12%, respectively. Similarly, the neural network-based models (**RNNPS**: 82.66%, **GRUPS**: 76.14%, **CNNPS**: 82.55%) return relatively higher mean TPR than the traditional ML models. But PResUP outperforms all these baselines with a mean TPR of 86.07%. These findings demonstrate that - (a) aggregating data from all users without the physiological response similarity (as done in **GPS**), or (b) combining data just based on demographic similarity (as done in **GBPS,** and **APS**) are not good choices. Rather, aggregating data from users having similar physiological responses (as done in PResUP) helps to obtain superior performance.

Finally, we note that **TBS** has the best mean TPR (100%). This is because it continuously probes at every 5-seconds interval, therefore all opportune probing moments are captured but at the cost of very high probing rate (Section VI-B) and very high error rate (Section VI-C2).

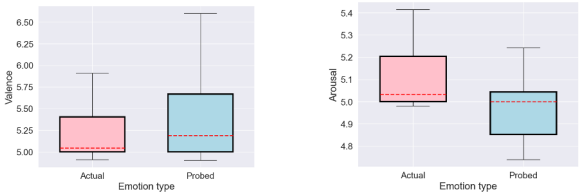
2) Error Analysis: In this section, we analyze the error rate in detecting the opportune probing moments (see Table I). We note that the **TBS** model performs the worst (mean FPR: 100%) as it does the probing continuously and, therefore, ends up probing in every inopportune moment. The **PPS, FBS,** and **GPS** baselines also have a high mean FPR of 16.16%, 15.02%, and 14.93%, respectively. The demographic-based model, **APS, GBPS** also have a high mean of FPR of 14.35% and 13.68%, respectively. Similerly, the neural network-based models (**RNNPS**: 12.88%, **GRUPS**: 13.68%, **CNNPS**: 15.47%) also have high mean FPR. On the contrary, PResUP outperforms all the baselines with a mean FPR of 12.27%. These findings further underscore the effectiveness of PResUP that not only it detects the opportune moments correctly (as TPR is high) but also makes relatively few errors while detecting the opportune probing moments.

D. Trade-off: Inopportune Moments and Annotation Cost

We performed trade-off analysis between disregarding inopportune moments and continuous annotation effort. Since the inopportune segments can't be completely discarded to train the opportune moment detection model, we investigated the impact on annotation effort by varying inopportune segments from 10% (very few) to 100% (i.e., considering all inopportune segments). Initially, the probing rate was as high as 15.24 (for 10% inopportune segments), which reduced to 7.12 (for 50% inopportune segments) before returning the final probing rate of 5.8 (using all inopportune segments). This is because with more inopportune segments, the model learns effectively to distinguish between opportune and inopportune segments. We also observe that the large reduction in probing rate (≈62%) does not influence the TPR (≈5% change) and FPR (≈1% change) substantially.

E. Emotion Annotation Quality Comparison

We investigate the emotion annotation quality by comparing the user-wise valence (and arousal) sampled using PResUP (at opportune moments) and present in original continuous annotations. To compare the users' valence (and arousal), we perform the following steps (once for valence and once for arousal). We compute the median score of valence for every user by collecting annotations only at those windows, when the probe was issued as per the PResUP framework. Similarly, we compute the user-wise median score of valence from the ground truth data. Then, we checked if the valence (actual and sampled) and arousal (actual and sampled) follow normal distribution using the Shapiro-Wilk test [33]. This experiment revealed that the responses did not follow a normal distribution ($p < 0.05$) both for valence and arousal. Since the distribution is not normal and the two groups (actual, sampled) are not paired, we performed Mann-Whitney U test [34] to evaluate the difference between actual and sampled valence scores. The same steps were repeated for arousal. We also compared the video-wise valence (and arousal) sampled using PResUP (at opportune moments) with the original continuous annotations and found no significant differences (see supplementary file).



(a) Comparison of valence scores **(b)** Comparison of arousal scores

Fig. 5: Mann-Whitney U test reveals no significant ($p < 0.05$) difference between the continuous and probed valence (and arousal).

The comparison for the valence and arousal are shown in Fig. 5a, and 5b respectively. The medians of valence (actual) and valence (probed) are 5.046 and 5.189, respectively. The Mann-Whitney's U test did not find a significant difference in the actual and probed values (U = 392.5, Z = 0.863, p = 0.387, r = 0.11). Similarly, the medians of arousal (actual) and arousal (probed) are 5.032 and 5.0, respectively. In this case also, we did not find a significant difference in the actual and probed values (U = 563.5, Z = 1.701, p = 0.090, r = 0.22) from the Mann-Whitney's U test. In summary, these findings underscore that there is no significant difference in the valence and arousal values between the ground truth continuous annotations and the values sampled using the PResUP framework.

VII. Generalization of PResUP: K-emocon Dataset

We evaluated the generalizability of PResUP on the publicly available K-emocon [8] dataset. This is an audio-visual dataset that records physiological responses from three wearable devices during 16 paired debates (a total of 32 participants) on a social issue. The continuous annotations were collected on a 5-point valence-arousal scale as per the Circumplex model [17] of the emotion. We implemented the pre-processing steps outlined in Section III-A, which yielded a total of 6644 segments (13.6% opportune, 86.3% inopportune), thus on average 238 segments per user. We applied PResUP on this processed dataset to detect the opportune moments.

	Probing rate↓	**TPR (%) ↑**	**FPR (%) ↓**	**LR+ ↑**
TBS	237.28 (0.00)	100 (0.00)	100 (0.00)	1.00
PPS	74.46 (14.87)	81.26 (15.41)	13.01 (1.02)	6.25
FBS	32.32 (16.61)	80.22 (16.19)	4.04 (5.86)	19.86
GPS	33.89 (21.96)	75.65 (36.17)	5.16 (6.58)	14.66
APS	33.82 (16.41)	78.28 (17.51)	3.72 (5.74)	21.04
GBPS	33.78 (17.22)	80.01 (17.16)	3.74 (5.98)	21.39
RNNPS	34.35 (23.93)	68.39 (44.94)	6.56 (6.92)	10.42
GRUPS	40.96 (18.35)	50.51 (43.57)	11.60 (2.94)	4.35
CNNPS	44.5 (16.38)	73.53 (27.01)	11.41 (5.95)	6.44
PResUP	**31.36 (11.23)**	**82.26 (13.38)**	**3.41 (4.41)**	**24.12**

TABLE II: Performance comparison of PResUP and baselines on K-emocon dataset. The values indicate the user-wise average and std.dev (inside parenthesis) for a metric. PResUP outperforms all baselines in terms of probing rate (least probing rate), LR+ (highest LR+), TPR, and FPR. Although **TBS** has the highest TPR, it also has the worst FPR (100%). ↑, ↓ indicate higher and lower value preferred respectively.

The performance comparison of PResUP and the baselines on the K-emocon dataset is presented in Table II. We note that PResUP outperforms all the baselines in terms of probing rate (31.36), therefore offers an average reduction of 25.71% in probing rate compared to the baselines. It also has the highest LR+ (24.12), highest mean TPR (82.26%), and the least mean FPR (3.41%). We also do not observe a significant difference in the annotation quality for valence and arousal (results presented in supplementary file). These findings showcase that PResUP reduces the continuous emotion annotation effort without compromising the annotation quality.

VIII. Conclusion

This paper proposes PResUP, a framework to collect the emotion self-reports opportunistically so that continuous emotion annotation effort is reduced in video platforms. To achieve this, first, it constructs the physiological response profile of the users and clusters the users based on the profile similarity. Later, combining physiological response data among the users present in a cluster, an LSTM model is constructed to detect the opportune probing moments for emotion self-report collection. We validated the proposed approach on a publicly available continuous emotion annotation dataset (CASE). The major findings from the evaluation are that PResUP reduces the probing rate by 30.3% (on average), detects the opportune probing moments with a TPR of 86.1% (on average), and yet maintains the annotation quality as collected during continuous annotation. These findings generalize on another public dataset (K-emocon) highlighting the effectiveness of PResUP to improve the continuous emotion annotation in video platforms.

IX. Acknowledgement

This research has been supported by the CDRF grant (C1/23/152) of BITS Pilani Goa, Chanakya Ph.D. Fellowship at AI4ICPS Innovation Hub (IIT Kharagpur), and the SURE grant (SUR/2022/001965) of SERB (Science and Engineering Research Board) of the Department of Science & Technology (DST), Government of India.

Ethical Impact Statement

This work aims to reduce the continuous annotation effort leveraging physiologocal responses. To acheive this, a machine learning model is developed that identifies opportune moments of emotion self-report collection (rather than continuous annotation). We use two publicly available datasets CASE and K-emocon to train and validate the model. The proposed framework is non-obtrusive, uses the physiological response data (heart rate, GSR) as available in these datasets. The possible applications of this technology are any video-based applications (e.g., online gaming, tutoring) that utilize emotional cue to improve user engagement.

Although we aim to develop emotion annotation tool to reduce the continuous annotation effort, we understand that capturing emotion self-reports has the potential to express sensitive details about a user's mental state, posing risks of misuse or exploitation. Finally, we also recognize the possibility of inherent bias in our models as they are trained on datasets collected from individuals belonging to a geographic location or culture.

References

[1] P. Murali, J. Hernandez, D. McDuff, K. Rowan, J. Suh, and M. Czerwinski, "Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[2] M. A. Hasan, N. F. M. Noor, S. S. B. A. Rahman, and M. M. Rahman, "The transition from intelligent to affective tutoring system: a review and open issues," *IEEE Access*, vol. 8, pp. 204 612–204 638, 2020.

[3] Y. Gao, Y. Jin, S. Choi, J. Li, J. Pan, L. Shu, C. Zhou, and Z. Jin, "Sonicface: Tracking facial expressions using a commodity microphone array," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–33, 2021.

[4] 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," in *ACM CHI*, 2020, pp. 1–15.

[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 *ITRW Speech-Emotion*, 2000.

[6] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, 2017.

[7] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data*, vol. 6, no. 1, 2019.

[8] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, pp. 1–16, 2020.

[9] P. Verduyn and S. Lavrijsen, "Which emotions last longest and why: The role of event importance and rumination," *Motivation and Emotion*, vol. 39, no. 1, pp. 119–127, 2015.

[10] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[11] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[12] S. Lin, R. Clark, R. Birke, S. Schönborn, N. Trigoni, and S. Roberts, "Anomaly detection for time series using vae-lstm hybrid model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2020, pp. 4322–4326.

[13] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.

[14] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[15] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.

[16] J. E. LeDoux and S. G. Hofmann, "The subjective experience of emotion: a fearful view," *Current Opinion in Behavioral Sciences*, vol. 19, pp. 67–72, 2018.

[17] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[18] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *2013 humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 709–710.

[19] J. M. Girard and A. G. Wright, "Darma: Software for dual axis rating and media annotation," *Behavior research methods*, vol. 50, no. 3.

[20] M. A. Islam, M. S. H. Mukta, P. Olivier, and M. M. Rahman, "Comprehensive guidelines for emotion annotation," in *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, 2022, pp. 1–8.

[21] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 857–869, 2019.

[22] S. Yang and G. Yang, "Emotion recognition of emg based on improved lm bp neural network and svm." *J. Softw.*, vol. 6, no. 8, pp. 1529–1536, 2011.

[23] S. L. Joshi, R. A. Vatti, and R. V. Tornekar, "A survey on ecg signal denoising techniques," in *2013 International Conference on Communication Systems and Network Technologies*. IEEE, 2013, pp. 60–64.

[24] A. Adithya, S. Tiwari, S. Sen, S. Chakraborty, and S. Ghosh, "Ocean: Towards developing an opportunistic continuous emotion annotation framework," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 9–12.

[25] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[26] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with granger causality and dynamic causal modelling," *Current opinion in neurobiology*, vol. 23, no. 2, pp. 172–178, 2013.

[27] A. Wismüller, A. M. Dsouza, M. A. Vosoughi, and A. Abidin, "Large-scale nonlinear granger causality for inferring directed dependence from short multivariate time-series data," *Scientific reports*, vol. 11, no. 1, p. 7817, 2021.

[28] A. Swain, V. Ganatra, S. Saha, A. Mathur, and R. Phadke, "P-lstm: A novel lstm architecture for glucose level prediction problem," in *International Conference on Neural Information Processing*. Springer, 2022, pp. 369–380.

[29] "Granger causality tests," 2024, https://tinyurl.com/funzzup7.

[30] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[31] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[32] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–21, 2018.

[33] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[34] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.