Choose or Fuse: Enriching Data Views with Multi-label Emotion Dynamics

Xi Laura Cang cang@cs.ubc.ca
Rubia Reis Guerra rubiarg@cs.ubc.ca
Paul Bucci pbucci@cs.ubc.ca
Bereket Guta bguta@cs.ubc.ca
Karon MacLean maclean@cs.ubc.ca

Laura Rodgers, Hailey Mah, Shinmin Hsu, Qianqian Feng, Chuxuan Zhang, Anushka Agrawal
Department of Computer Science
University of British Columbia
Vancouver, Canada

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953882

***Abstract*—Many emotion classification and prediction approaches focus on emotion *state*, defined as static and single-valued. In contrast, our in-body experience is of sensations that can quickly evolve, consistent with scientific evidence of physiological regulation mechanisms. Can we reframe classification to estimate dynamic emotion parameters at interactive rates?**

For insight into dynamic emotion characteristics, we developed a multipass labelling protocol to capture controlled yet genuine emotion evolution elicited as 16 participants played a tense video game. We analyze and align multiple self-report outputs, inspect the signals for emotion dynamics, and consider label metaphors of position and angle – "where I am" *vs.* "where I'm going". Finally, we reflect on the benefits and drawbacks of such a protocol for developing models of fast-evolving emotion.

***Index Terms*—affective computing, emotion classification, emotion labelling, emotion dynamics**

I. Introduction

Whether building robots that detect anxiety through touch interaction or video games that dynamically adjust level difficulty to optimize player engagement, computational models of authentically developing emotions are the foundation of technology. Challenges arise in developing these computational models from true and spontaneously evolving emotions.

Emotion theorists have long observed time-varying dynamics of emotion expression, attributing them to complex neurological and physiological regulation mechanisms [1], appraisal effects [2], cognition and contextual factors [3], [4]. To simplify in-lab research, computational emotion modelling often relies on an "emotions-as-point" metaphor [5], [6], represented as a dimensionless point in an emotion plane in which self-reporting static emotion labels for classification involves easy-to-read scales, often along dimensions of arousal, valence, and dominance [7]. While these models are convenient, for realtime use we need to recognize emotion evolution over time, rather than distilling a lengthy event into a single label.

Going from theoretical to computable: Obtaining authentic emotion data is a significant obstacle. Our memories and emotional assessments are affected by time and reflection [2], [4]; how representative can a reporting scheme be of someone's "reality"? Commonly used labels on the arousal-valence circumplex model [8] or PANAS [9] or SAM [10] (among others) quickly become intractable for sampling at the rates in which emotion can potentially evolve.

Emotion is personal: Independent of the measurement instrument, self-report of emotion incites questions of generalizability across the population. A researcher's understanding of the instrument scale may be very different from that of a participant [5]; our comprehension of an emotional 'landscape' or internalized emotion frames of reference are highly subjective, influenced by life experiences and personal history [11]. We presume that any set of ground-truth labels for self-reported emotion are similarly personalized: *i.e.*, the experience or scale for *anger* for one person may not be recognizable for another.

We propose that evaluating emotion based on dynamic qualities will advance the accuracy of machine recognition of human emotion experiences. Better forecasting of a user's near-future emotional expression allows for system responses that are temporally and situationally appropriate.

A. Approach

We assess the viability of building computational emotion models based on **dynamic** conceptualizations of emotion change to bolster our capacity to predict and respond to human emotion based on observed behavior or self-reports. Multi-pass labelling requires high investment in early model building, which can pay off by highlighting how to optimize labeling in later real-time use. As outlined in Figure 1, this paper evaluates reporting consistency between passes of a data collection and emotion labelling methodology, leaving model building and classification performance for future work.

Specifically, we reflect on our multi-pass protocol which (a) **triangulates** emotion self-reports with modality-agnostic observable data; and (b) employs co-creation of **personalized** calibrated emotion scales which form the frame of reference for multi-pass self-reports, collected with minimal intrusion on the primary emotion event. Using a joystick for spatiotemporally high-resolution post-hoc ratings, we can construct data windows that are (c) **versatile** to accommodate a variety of emotion metaphors at our choice of time scale.

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC; grant RGPIN-2018-04828) for supporting this work.

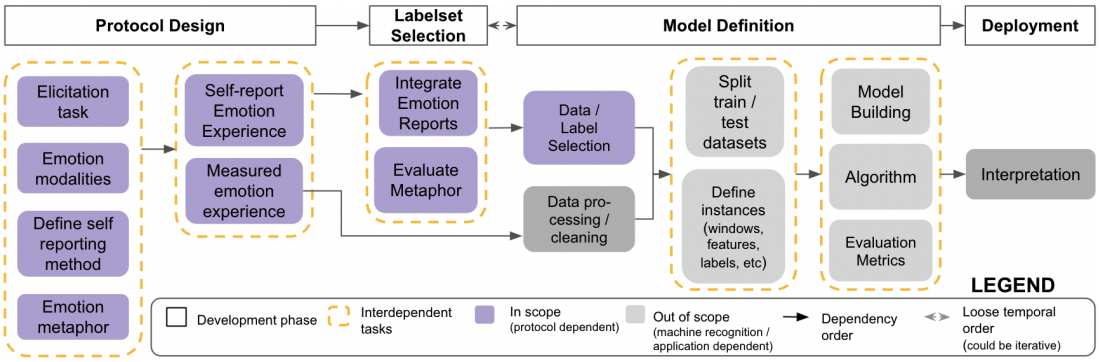


Figure 1: Roadmap for developing an emotion-prediction engine for an emotionally responsive application.

B. Research Questions

Two lines of inquiry guided our assessment of this approach to operationalize real-time models for emotion forecasting.

1. Do user-centered emotion reports add new information? Nuances in users' emotion language, manifesting as apparent inconsistency, can interfere with emotion model performance and validity [12]. We center users by including (a) personally calibrated emotion scales where we create a shared understanding of instruments and measures [5], and (b) multiple labelling passes at different resolutions and retrospective distance; then assess the information gained from these elements. For example, do people rank common emotion words similarly? In what ways does labelling data differ by pass? What do we gain from quantifying the differences?

2. How might we incorporate the dynamic nature of emotions into our computational models? Operationalizing dynamic emotions requires models that represent the natural evolution of an emotional experience.

We begin with the prevalent movement-based metaphor of *emotions-as-position* ('where I am', an ordinal value on an emotion scale), and propose another of *emotions-as-angle* ('where I'm going', the direction and sharpness of change). We add to these previously-proposed emotion dynamic measures of inertia, instability, and variability [13], and compare the properties of each with each other and in between-participant variability for insights into how they might have value for responsive computational models.

Through these investigations, we contribute:

- A multipass labelling protocol with insights into how to employ triangulated emotion labels, including the role of personalized emotion word calibration;
- Insights into the descriptive properties of various dynamic emotion parameters, relating to their potential for use in responsive computed models.

In the following, we root our protocol development in the existing literature, describe the devices and instruments we created to measure continuous dynamic emotion, outline the data collection procedure, and evaluate the data according to our questions. In discussing our findings, we consider where these new model elements may provide the greatest value.

II. Related Work

Protocols featuring internally consistent emotion metaphors, measurement instruments, and elicitation procedures increase the likelihood of representing true participant experiences [5].

A. Emotion Self-Report

Classifying emotion requires capturing and labelling emotional experiences. Representation thus impacts how we ask users to report their experience.

Russell's circumplex model [8] is a commonly used instrument depicted as a spatially continuous 2D space of arousal and valence (plus dominance in 3D [14]). It underlies popular labelling schemes, most involving a participant locating emotion words on its axes; *e.g.,* words associated with PANAS, the Positive-Negative Affect Schedule [9].

The Self-Assessment Manikin (SAM [10]) makes this more natural with Likert scale dimensions [15], [16].

Natural language reporting methods are used when experiences (maybe a self-contained memory [17], or a touch [18]) are sufficiently brief, simple to fit a single label, and precede an opportunity for the participant to report without experiential interference. They become intractable for segments that are longer than a few moments, span multiple emotions, and/or require rapid computed response (before the segment ends).

Still with a dimensional representation, others have collected *temporally continuous* emotion ratings using a mouse- [19] or a joystick [20], [21]. For hands-free activities, a joystick allows for high temporal-resolution concurrent reporting, but at the cost of emotional intrusiveness. Post-hoc ratings require review of a recorded experience.

We drew on these approaches to design our own **joystick-based continuous emotion annotation** system.

B. Characteristics of Emotion Dynamics

The methods above imply emotion as "state". Even models that feature sequences (i.e. Bayesian emotion models) denote each stage as a single state [4], [22].

Regarding emotion instead as a *process*, as in appraisal theory [2], may better reflect human experience; but this perspective must be operationalized. One approach is to calculate *emotion dynamics*, by quantifying progression in three fluctuation parameters on one's emotional movement:

(1) *inertia* (the time it takes), (2) *instability* (by how much), and (3) *variability* (the range of those changes), calculated as autocorrelation, mean square of successive differences, and within-subject variance respectively [13], [23]. Using these summary metrics over a report time series, researchers have evaluated emotional character arcs in movies [24], examined the role of exercise in emotion regulation [25], and even predicted mood disorders [23].

Can we use these markers at high resolution, to capture transitions and support concurrent response or are other motion characteristics more appropriate? We investigate **sourcing labels from a report's emotion dynamics.**

C. Labelling and Timing

Timing is key to regulation, reflection, reporting, and in-event reactivity. Emotions evolve at multiple time scales; an event may evoke a different emotion after cognitive reflection on an in-time reaction [4]. The optimal timing for capturing a self-report is complex. Too soon may curtail rich and valuable reflection [2]; too late incurs memory decay [26]. Concurrent emotion evaluation is typically impractical: probing for labels is intrusive and distracting – naming a feeling is a form of reflection and regulation [1], [27], [28].

To capture reflection and generate training data for future responsive models, **we collect reports in two passes and use multi-timescale labelling** – giving time for self-reflection, and mitigating memory degradation with video reminders.

D. Emotion Elicitation

Where applications require in-time recognition of emotion, data must represent realistic emotion expression [29], [30]. Relived or recalled emotion is one proxy [7], [31]. Participants are prompted with an emotion word (the single label) and asked to recount the story of a past intense experience.

While successful in eliciting authentic and wide-ranging responses, this oversimplifies an episode to emotive homogeneity [7]. Furthermore, participant stories are hyper-individualistic, not amenable to a search for commonalities. Conversely, entertainment media can root participants in a more uniform elicitation stimulus, with many validated video and music clips used successfully for this purpose [27], [32]. Video games have shown promise in producing physiological responses analogous to that of real life evocations [33].

Here, **we use a horror video game to elicit emotion**. This genre has shown high user immersion and engagement, evoking emotions from anxiety to happiness and contentment [34].

III. Data Collection Protocol

Our priority was to obtain triangulating data views on the emotional space of momentary transitional experiences. $N = 16$ individuals (8 reporting as male, 8 female; 19 to 34 years of age, half under 25) participated. Each participant supplied self-report data that demonstrates our protocol, by completing four tasks as outlined in Figure 2 and detailed below.

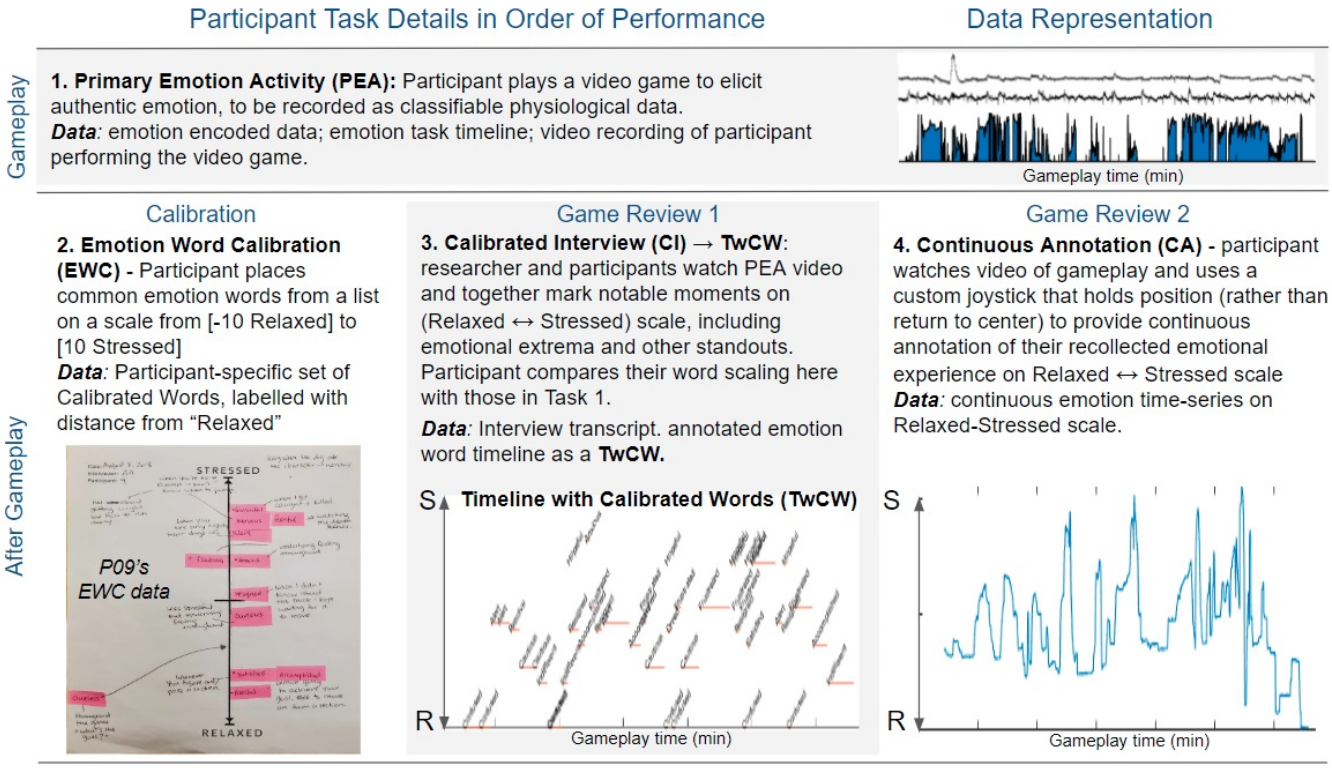


Figure 2: Participant tasks and resulting data. At lower left is an EWC example: word stickers placed on a Relaxed-Stressed scale, plus P09's other annotations. The latter resulted from P09 later contextualizing their in-game experience.

A. Participant Task 1: ***Primary Emotion Activity (PEA)***

To demonstrate this protocol, we use video game play to elicit authentic and spontaneous emotion. We chose *Inside* [35] to stimulate anxiety without graphic violence and spark moments of accomplishment or satisfaction, all with easy-to-learn keyboard controls. We selected participants for their affinity for video games, excluding those with experience of *Inside*.

For reviewing the primary gameplay experience in later passes, we videorecorded participants' faces and game screen (OBS[1], 30fps). Gameplay averaged 13:24m (min 8:25, max 21:37, SD 3:88).

B. Participant Task 2: ***Emotion Word Calibration (EWC)***

To contextualize individual interpretations in later steps, participants rated up to 15 emotion words, two write-ins and 13 from the PANAS [9]: *Cautious, Satisfied, Hopeful, Frustrated, Anxious, Nervous, Threatened, Resigned, Alert, Accomplished, Fearful, Dread, Curious*. Figure 2, lower left shows P09's sample scale, ordering these words between Relaxed to Stressed (chosen to represent diametrically opposing quadrants from Russell's circumplex of Arousal *vs.* Valence [8]).

We measured the distance from the Relaxed line to each word's placement, mapped it to a 20-point scale ([-10,10]), and aligned the words and their scaled heights with the interview (I) transcript via timestamps, to form a time-series of emotion word (and synonym) height.

C. Participant and Researcher Task 3: ***Calibrated Interview → Timeline with Calibrated Words***

In the first labelling pass, participant and researcher jointly reviewed the gameplay video. The participant indicated emotionally notable points while the researcher marked them on a gameplay timeline. Because participants had previously undergone a word calibration, they were primed to consider how the offered vocabulary were distributed across the emotion scale.

From the Task 2 Interview transcript, we found synonyms and root words using Python's Natural Language Toolkit [36]. We constructed the **Timeline with Calibrated Words (TwCW)** by placing values where a root matched the EWC, with each value a numerical distance from Relaxed. For example, P09's comment *"The barking in the distance filled me with anxiety"* would map the calibrated point value of 14-Anxious (synonym of *Anxiety*) on P09's calibrated 20-pt scale at the timestamp in the game where the dogs began barking.

Participant language included μ=37(σ=7) calibrated word instances with annotation frequency μ=0.05(σ=0.015) words/min; duration was roughly double gameplay.

D. Participant Task 4: ***Continuous Annotation (CA)***

In the final pass, participants reviewed the PEA video without pause. They used a custom joystick (holds position rather than returning to center) to continuously trace a 1-dimensional emotion rating between predefined extremes (inspired by [19]–[21]), here employing the previously calibrated axis. The result is a continuous rating time-series (256Hz) corresponding to the original gameplay, downsampled for analysis to 30Hz to match the video framerate.

During annotation, smoothed joystick position is graphically rendered as the height of a bar on the video screen, for feedback on proximity to a more Relaxed (blue) or Stressed (pink) emotional moment.

E. Task Order

Task order was carefully chosen to minimize influence on emotion elicitation while increasing the likelihood that participants would use a common set of emotion words to describe their experience. During Step 3, the interview allowed players to explicitly process their emotions out loud, guided by researchers looking for notable emotional events – strong emotions, startling or uncomfortable moments, odd behaviour etc. Leaving the joystick evaluation as the final step lets participants internalize and contextualize the emotion scale in preparation for the continuous annotation.

IV. Exploring Multi-Pass Emotion Self-Reports

Our present analytical goal is to explore the properties of and relationships among the reports obtained with this protocol, primarily by examining the degree and nature of their [dis]similarity over a range of metrics, and probing for physical intuition among them.

A. Commonality in Interpreting ***Emotion Words***

To assess across-participant similarity of calibration ratings (as a proxy for model generalizability), Figure 3 plots *rating variance* for each of the calibrated words in order of decreasing agreement (increasing variance).

For a quantitative view of cross-participant consistency, we also conducted an *intra-class correlation (ICC)* (inter-rater reliability test [37]). For the subset of emotion labels rated by all participants (*Anxious, Cautious, Frustrated* and *Satisfied*), we found ICC(2, k=16)=0.99, $p \ll 0.01$ ($\alpha = 0.05$, $CI = [0.97, 1.0]$), based on mean rating over an absolute-agreement, 2-way random-effects model. ICC values > 0.9 indicate high reliability [37], suggesting these ratings are overall highly similar across-participant for this set of emotions. Indeed, the four rated by all participants had an ICC(2, k=16) of 0.99.

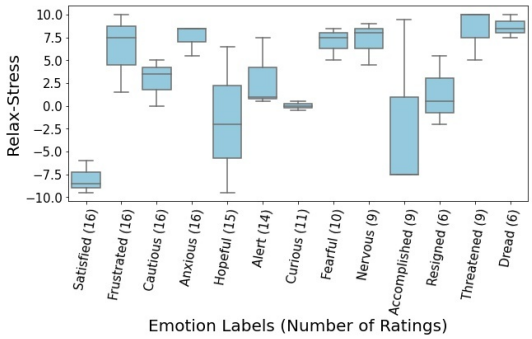


Figure 3: Rating variance by calibration word, ordered by number of participants who provided a rating for that word.

[1]Open source video recording and streaming. https://obsproject.com/

However, this agreement varies as set size increases, first decreasing monotonically then dropping sharply at *Satisfied - Resigned* to ICC(2, $k = 4$)=0.83. This may be partially due to the relative sparseness of ratings.

Taken together, these results support that there are **substantive differences in how individuals interpret emotion words, highlighting the importance of personalized models.**

*B. Self-Report Modality Consistency via **Time Series***

High similarity between self reports indicates consistency and perhaps interchangeability of report modalities; differences might suggest invalidity of one or both, or that they capture different information. Interpreting within-participant TwCW and CA as time-series, we use standard time-series analysis methods [38] (with appropriate condition verification steps) to check for signal similarity – Pearson's correlation – and confirm that both data streams are appropriate responses to a common stimulus – Granger's Causality [39]).

Test Preparation: Using raw report data, we first confirmed that **both time-series were stationary** with the Augmented Dickey-Fuller (ADF) test (Bonferroni-Holm correction $\alpha = 0.05$, $p_{\mathrm{BH}} < 0.02$[2]), and that their statistical properties did not change over time [40]. Prior to evaluating cross-correlation between the two reports, we verified that each was not auto-correlated to avoid artificially inflated correlations [41]). With Python's *statsmodels* [42], all peaks were at lag=0 for all participants' TwCW and CA auto-correlation plots (*i.e.*, both signals present low correlations at all lagged versions of itself). We conclude that **neither signal is self-similar**.

The TwCW and CA self-reports are sampled at different times and resolutions (0.05Hz and 30Hz respectively). We downsampled the CA series rather than interpolate the sparse TwCW, to minimize bias.

Pearson's Correlation for signal similarity:[3] P01, P02, P08, and P14 had moderate correlation coefficients for the two emotion self-reports (CA and TwCW) at $\rho > 0.3$ ($p_{\mathrm{BH}} < 0.05$). However, in general there was no significant correlation between the report streams: p-values exceed the threshold after a Bonferroni-Holm's adjustment to $\alpha = 0.003$. We infer that **individuals' self-reports differed** in the metrics we observed.

Granger Causality Test for source plausibility: Although Granger cannot confirm direct causality between different variables [43] (*i.e.*, it does not claim TwCW causes the CA values), we employ the test to evaluate whether time-series for CA could *forecast* TwCW and vice versa. We employed a Bonferroni-Holm correction ($\alpha_{\mathrm{BH}} = 0.05/N$, N = number of participants). We found significance for 15 of 16 participants ($p_{\mathrm{BH}} < 0.048$), suggesting that one label stream could be used to forecast the other for all except P02. This implies **the data streams are appropriate as responses to the same stimulus**.

[2]For all except P01 (TwCW): $p_{\mathrm{BH}} = 0.07$, ADF test statistic $= -2.671$

[3]Pearson's correlation results at $\alpha = 0.05$: P01 ($\rho = 0.38$, $p_{\mathrm{BH}} = 0.142$), P02 ($\rho = 0.38$, $p_{\mathrm{BH}} = 0.235$), P08 ($\rho = 0.43$, $p_{\mathrm{BH}} = 0.235$), P14 ($\rho = 0.37$, $p_{\mathrm{BH}} = 0.245$)

C. Comparing Motion Characteristics of Emotion Dynamics

We next examined how various parameters computed on these time series might reveal differing insights. In this scope we included: signal *Position* (the prevalent standard, and following an "emotion-as-state" metaphor); *Angle* (drawing on an alternative metaphor for emotion as directional and changing); and [13]'s three emotion dynamic parameters of *Inertia*, *Instability* and *Variability*. Our investigation included comparing these time series (original and computed) through summary statistics and histograms, all by participant.

Data Preparation: We further analyzed each participant's Continuous Annotation[4]data by first partitioning the continuous self-report data into 500ms windows (window count μ=1587.75, σ=462.50 by participant). Where window boundaries do not coincide with a logged data point, we imputed with the previous data point, turning our time-series into a higher-resolution but stepped signal.

We computed *Position* labels from windows by mean value; and *Angle* labels as the rate-of-change per minute from a least squares linear fit, in the form of an angle $\theta \in [-\pi/2, \pi/2]$. Using R's *psych* package [45], we calculated *Inertia* (autocorrelation coefficient), *Instability* (Root Mean Square of Successive Differences (RMSSD)) and *Variability* (Standard Deviation (SD)) by window for each participant [13].

Comparing Summary Statistics and Histograms by Parameter: Figure 4a shows signal statistics for each participant and parameter. The means for all five measures track closely across participants. However, spread differs: *Inertia* is relatively tight and symmetric, *Variability* is broad and highly asymmetric, *Instability* in between.

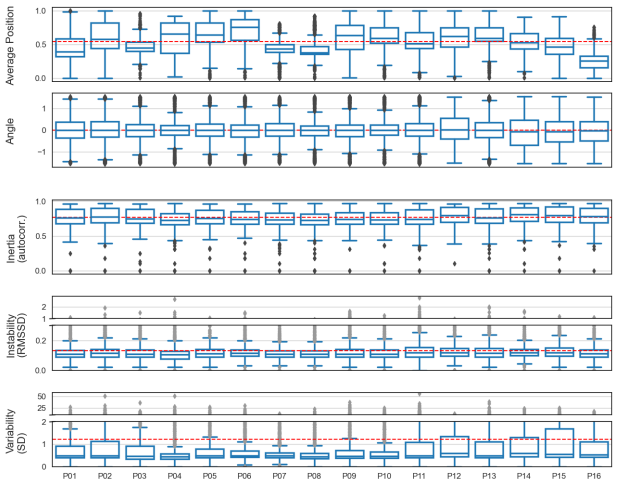
In an alternative view, Figure 4b shows the same parameters and signals, but now as histogram distributions. Data for these four participants are reasonably representative.

Comparing these two representations of the same underlying data is insightful. For example, while in Figure 4a *Position* is clearly less stationary than *Angle*, 4b indicates the form that this takes (broader spread, spikiness). And while the dominating feature of the other three ED's boxplots is the uniformity of means across participants, histograms reveal their internal parameters as starkly different: *Inertia* is broad and high-valued, the others low-valued with very long tails.

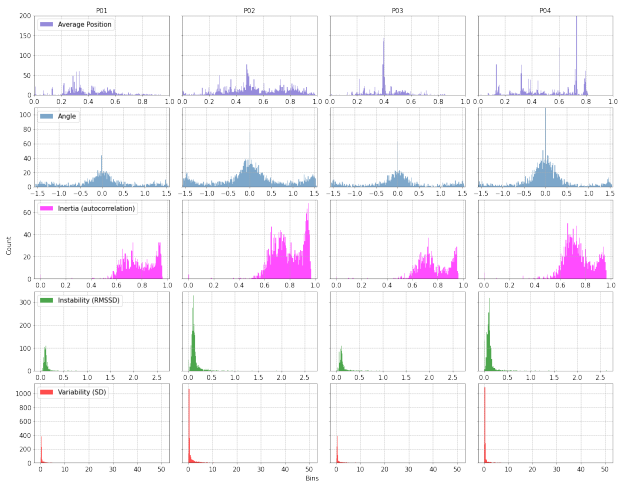
No insight was gained from visual analysis of spectral qualities (from a Fast Fourier Transform) of all five parameters.

Which is Best? The preceding section's results demonstrate that the relatively high resolution of the CA report (30Hz raw, parameters computed at 2Hz) affords computation of a variety of descriptive parameters. Getting to the root of what the differences in label representation mean will require approaches assisted by synchronized physiological data views.

[4]Tests for equivalence between the two sets of self-report (CA and TwCW) across each of the three emotion dynamics parameters (two 1-tailed paired samples t-tests [44] per dynamic measure) were inconclusive ($p > 0.5$, $t(15) \ll 0.001$, $d \ll 0.001$). Subsequent emotion dynamics explorations were done on the higher resolution CA data.



(a) Boxplots of emotion dynamics of Continuous Annotation (Task 4) data, by Participant ($N = 16$). Position ($M = 0.5465$, $SD = 0.2221$), Angle (0.0049, 0.7127), Inertia (0.7666, 0.1215), Instability (0.1316, 0.1086) and Variability (1.2165, 2.4079).



(b) Representative subset of label distributions: emotions-as-position (average position; *purple*), emotions-as-angle (angle; *blue*), Inertia (*magenta*), Instability (*green*), Variability (*red*). Note that longer gameplay results in more samples.

Figure 4: Comparison of summary statistics and histograms by emotion parameter.

V. Discussion

Compared to past studies of dynamic changes in behaviour or mood [13], our video game task is short and densely reported. With its data we reflect on our questions and protocol, highlighting implications for high-resolution real-time models.

A. Multi-Pass and Personalized Emotion Reporting

To estimate emotion evolution by-the-second, we can select a single dimensional emotion scale and collect self-reports (as in our CA data). How does adding scale calibration and a review/interview phase enrich this report stream?

Personalized scales clarify what may be generalizable, as well as improving personal models' accuracy: Asking participants to project a set of emotions onto a specified emotion axis grounds the ratings in an individualized experience between the Stressed-Relaxed extremes. Plotting the ratings across commonly used words (as in Figure 3), we see that words with low rating variation – *Satisfied* and *Anxious* – may be useful as emotion reference frames. In contrast, high variance words like *Hopeful* or *Accomplished* may be less useful for labelling without additional interpretation.

Multipass reporting increases label versatility: A continuous annotation of emotion communicates a highly personal experience at a resolution that is otherwise difficult to solicit. As a continuous quantitative signal, we can model emotion as a regression for high-resolution forecasting or elect to discretize (or bin values) for categorical classification. Additionally, we can compose an entirely new time-series by incorporating our personalized scale into an interview as a lower resolution signal where continuous annotation is impractical or unnecessary.

Disagreement may indicate synergy, not conflict: Data from our two passes (annotation and interview) are not correlated enough to be interchangeable, yet causality results indicate they are highly related. Perhaps each has its own authenticity and value, which could be optimized in protocol refinement, then extracted and integrated. Further work is needed to identify the different perspective that each brings.

B. Incorporating Dynamics into Emotion Models

Reading signal characteristics (like autocorrelation, mean successive differences, variance) as measures of emotion inertia, instability and variability connects them to lived experience. What can they mean for intuitive predictive models?

Momentary emotion dynamics as characteristic, not label: Inertia, instability, and variability can help elucidate "slow emotion" in mood disorders [23], but lose meaning in rapid-response timescales, and thus as emotion labels. Reframed as informative signal statistics, they yield hints such as emotion variability's larger spread suggesting extra *sensitivity* (Figure. 4a) which could inform model development, *e.g.,* by identifying archetypal behaviours for improved model selection.

An abundance of metaphors to fit the need: The metaphor of "emotion-as-position" does not capture "fast" emotion dynamics. For example, *Angle*, which captures relative differences in emotional intensity, has a natural physical meaning of directionality – *where I'm going*, not *where I am*. We have seen that *Inertia* and *Instability* respectively lend insight into responsiveness of emotion to stimuli, and emotive range.

Context may dictate choice of label metaphor. To identify if someone is *Excited*, we may choose a **position** representation; to catch *getting Sadder*, **angle** may work best. A **position** metaphor is more versatile; **angle** can be estimated from a set of points but the reverse requires additional information.

C. Protocol Reflections

At high temporal resolution, reporting can be intrusive and tedious. We reflect on our multipass labelling procedure for tradeoffs and consider possible improvements.

High-resolution labelling does not have to be intrusive. Since emotion reporting happens before and after elicitation, this labelling protocol accommodates any combination of sensing modalities. The emotion experience can unfold naturally, since labelling is done in review.

High time-resolution may be best for short time-scales. Continuous annotation is great for tracking emotion evolution during a 20-min video game session but onerous for prolonged review; and this protocol's overhead is unsuitable for occasional low-effort check-ins. Multiple passes are ideal for tasks that promote dynamic emotional experiences over a short time, and where reflection and review-dependent labelling are valuable: *e.g.,* therapeutic activities, recalling a memory, playing a game, interacting with an agent. Simultaneous emotion rating may be possible while watching a video or listening to music: joystick annotation during the elicitation, so long as the elicitation activity is hands-free.

Ordered tasks cannot be counterbalanced. We carefully selected the order of tasks to prioritize emotion reflection and recall. The tradeoff for lightening the mental effort and reducing time investment for multipass labelling means that we cannot counterbalance order for the Calibration (Step 2), Timeline with Calibrated Words (Step 3), and Continuous Annotation (Step 4). We are unable to evaluate generalizability of the labelling passes in other protocol orderings.

D. Future Directions

This paper is an initial exploration into the labelling procedure for dynamic emotion modelling. We highlight where future directions are highly promising.

Parameters computed on high-resolution data are different. What does this mean? To get behind different characteristics in computable descriptive parameters, one approach is to compare with other high-resolution data streams such as EEG and facial encoding. We plan to do this by focusing analysis on particular events (*e.g.,* timeline regions stimuli known to trigger reactions in all – a scary spot in the game), and see how these parameters look across multiple participants when calibrated in a variety of ways.

At what time scale does calibration change? We calibrated our scales prior to the emotion elicitation task. Could engaging in a highly emotionally charged activity influence the rating scale upon reflection? In future iterations of this protocol, we envision performing calibration tasks both at the beginning and end of the self-report labelling allowing us to investigate how calibration may drift within and between sessions.

How must models of dynamic emotion evolve? Longitudinal studies will reveal how to create personalized models that evolve with the individual. Mood, life and situational context influence perception of emotional events [46] but also change dramatically over time: we wonder how repeat data collection over the course of months impacts emotion models.

How to capture a range of emotion experiences? We selected a single-dimensional scale to simplify annotation; real-life events may trigger far more complex emotion landscapes where emotions are in conflict simultaneously (*e.g.,* feeling excited and sad about graduation). How can we make it more intuitive to document multiple simultaneous scales?

Choose or Fuse: Is report divergence an opportunity? Diverse self-reports may capture perspectives that are authentic in different ways. We have inspected characteristics of emotion self-report in the time- and frequency- domains.

Based on analysis insights, we might *choose* one approach, for its sensitivity or practicality. Or, we might *fuse* them, *e.g.,* using discrepant moments as a spotlight on emotional conflict or low-confidence labels. We plan to develop concrete choose-fuse strategies based on focused attribute study, which also lessen intrusion on emotion experience.

VI. Conclusion

We proposed a multipass data collection protocol to develop emotion models for real-time responsiveness in emotionally dynamic experiences. The protocol entails four sequential participant tasks: (1) emotion elicitation; (2) personal emotion calibration; and during video review, a (3) detailed interview and (4) continuous annotation of the emotion task. Using 16 participants' data, we determine that this multi-pass labeling implementation adds **versatility** to collection options, provides personalized and triangulated **insight into nuanced meanings**, and offers new options for **signal selection or integration**. We **show how emotion dynamics measures and metaphors can add value**, in particular *emotions-as-positions* or *-as-angles*; and propose promising next steps.

Ethical Impact Statement

We have proposed a novel multipass protocol for capturing and modeling high-resolution emotion experience at real-time scales. It is a personalization technique intended to benefit end-users: an automatable model evolution based on user input. While there is always potential for mal-use, this is mitigated by fundamental grounding in the individual rather than a generalized understanding of many. The investing user is the only beneficiary of model improvement; their data is of low value to others and less likely to invite exploitation.

Acknowledgments

We thank Dr Rebecca Todd and Dr James Kryklywy for the valuable insight into the neuropsychological effects of emotion evaluation that informed the design of this protocol. Many people have invested time and effort into this project: Kevin Chow, Tyler Malloy, Devyani McLaren, Andrew Moore, Drishtti Rawat, Zefan Sramek, Sherry Yuan, and Hafsa Zahid. This work has benefited significantly from their involvement. This work was funded in part by Natural Sciences and Engineering Research Council of Canada (NSERC) and conducted under UBC Ethics #H15-02611.

REFERENCES

[1] J. J. Gross, "Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology." *Personality & Social Psychology*, vol. 74, no. 1, p. 224, 1998.

[2] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, "Appraisal theories of emotion: State of the art and future development," *Emotion Review*, vol. 5, no. 2, pp. 119–124, 2013.

[3] B. Mesquita, L. F. Barrett, and E. R. Smith, *The mind in context*. Guilford Press, 2010.

[4] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[5] P. Bucci, X. Cang, H. Mah, L. Rodgers, and K. E. MacLean, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[6] P. Kuppens and P. Verduyn, "Emotion dynamics," *Current Opinion in Psychology*, vol. 17, pp. 22–26, 2017.

[7] X. L. Cang, P. Bucci, J. Rantala, and K. Maclean, "Discerning affect from touch and gaze during interaction with a robot pet," *Trans on Affective Computing*, no. 01, pp. 1–1, 2021.

[8] J. A. Russell, "A circumplex model of affect." *Personality & Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[9] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Personality & Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[10] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Behavior Therapy & Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[11] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, 2007.

[12] M. F. Jung, "Affective grounding in human-robot interaction," in *Int'l Conf on Human-Robot Interaction (HRI)*. IEEE, 2017, pp. 263–273.

[13] M. Houben, W. Van Den Noortgate, and P. Kuppens, "The relation between short-term emotion dynamics and psychological well-being: A meta-analysis." *Psychological bulletin*, vol. 141, no. 4, p. 901, 2015.

[14] I. Bakker, T. Van der Voordt, P. Vink, and J. De Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, 2014.

[15] T. Xie, M. Cao, and Z. Pan, "Applying self-assessment manikin (sam) to evaluate the affective arousal effects of vr games," in *Proc in Int'l Conf on Image & Graphics Processing*, 2020, pp. 134–138.

[16] A. Simoës-Perlant, C. Lemercier, C. Pêcher, and S. Benintendi-Medjaoued, "Mood self-assessment in children from the age of 7," *Europe's Journal of Psychology*, vol. 14, no. 3, p. 599, 2018.

[17] X. L. Cang, P. Bucci, A. Strang, J. Allen, K. MacLean, and H. S. Liu, "Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition," in *Proc of the 2015 ACM on Int'l Conf on Multimodal Interaction*, 2015, pp. 147–154.

[18] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka, "Touch communicates distinct emotions." *Emotion*, vol. 6, no. 3, p. 528, 2006.

[19] 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 *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[20] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *Trans on Affective Computing*, vol. 11, no. 1, pp. 78–84, 2020.

[21] T. Xue, S. Ghosh, G. Ding, A. El Ali, and P. Cesar, "Designing real-time, continuous emotion annotation techniques for 360 vr videos," in *Extended Abstracts of Int'l Conf on Human Factors in Computing Systems*, 2020, pp. 1–9.

[22] C. Conati, "Probabilistic assessment of user's emotions in educational games," *Applied Artificial Intelligence*, vol. 16, no. 7-8, pp. 555–575, 2002.

[23] S. H. Sperry, M. A. Walsh, and T. R. Kwapil, "Emotion dynamics concurrently and prospectively predict mood psychopathology," *Affective Disorders*, vol. 261, pp. 67–75, 2020.

[24] W. E. Hipson and S. M. Mohammad, "Emotion dynamics in movie dialogues," *Plos one*, vol. 16, no. 9, p. e0256153, 2021.

[25] E. E. Bernstein, J. E. Curtiss, G. W. Wu, P. J. Barreira, and R. J. McNally, "Exercise and emotion dynamics: An experience sampling study." *Emotion*, vol. 19, no. 4, p. 637, 2019.

[26] B. Dudzik and J. Broekens, "A valid self-report is never late, nor is it early: On considering the "right" temporal distance for assessing emotional experience." in *2nd Momentary Emotion Elicitation & Capture Workshop at CHI*, 2021.

[27] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[28] J. J. Gross, *Handbook of emotion regulation*. Guilford publications, 2013.

[29] Y. Gaffary, J.-C. Martin, and M. Ammi, "Haptic expression and perception of spontaneous stress," *Trans on Affective Computing*, vol. 11, no. 1, pp. 138–150, 2020.

[30] M. Hoque and R. W. Picard, "Acted vs. natural frustration and delight: Many people smile in natural frustration." in *IEEE Face & Gesture*, 2011, pp. 354–359.

[31] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[32] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment.* Oxford University Press, 2007.

[33] T. Terkildsen and G. Makransky, "Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence," *Int'l Journal of Human-Computer Studies*, vol. 126, pp. 64–80, 2019.

[34] F. Pallavicini, A. Ferrari, A. Pepe, G. Garcea, A. Zanacchi, and F. Mantovani, "Effectiveness of virtual reality survival horror games for the emotional elicitation: Preliminary insights using resident evil 7: Biohazard," in *Int'l Conf on Universal Access in Human-Computer Interaction*. Springer, 2018, pp. 87–101.

[35] D. Playdead, "Playdead's inside," https://playdead.com/games/inside/, 2022, accessed: 2022-04-21.

[36] Python, "Natural language toolkit - documentation," https://www.nltk.org/, 2022, accessed: 2022-04-21.

[37] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[38] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[39] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.

[40] W. A. Fuller, *Intro to statistical time series*. John Wiley & Sons, 2009.

[41] R. T. Dean and W. Dunsmuir, "Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models," *Behavior Research Methods*, vol. 48, no. 2, pp. 783–802, 2016.

[42] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[43] A. Shojaie and E. B. Fox, "Granger causality: A review and recent advances," *Annual Review of Statistics and Its Application*, vol. 9, pp. 289–319, 2022.

[44] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence testing for psychological research: A tutorial," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 2, pp. 259–269, 2018.

[45] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2022, r package version 2.2.3. [Online]. Available: https://CRAN.R-project.org/package=psych

[46] K. Hoemann, Z. Khan, M. J. Feldman, C. Nielson, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley, "Context-aware experience sampling reveals the scale of variation in affective experience," *Scientific reports*, vol. 10, no. 1, pp. 1–16, 2020.