What's In a Name: Issues in Triangulating Self-Reported Emotion to Establish Trustable Labels

Xi Laura Cang
cang@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

Qianqian Feng
fengqq@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

Karon E. MacLean
maclean@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

ABSTRACT

Multipass labelling is one way to add richness and reliability to self-reports of emotions. However, consideration of alternative data sources and perspectives of the same event stream exposes conflicts which must be resolved at analysis time in a way that provides an estimate of confidence and validity. As part of a larger project that is exploring the feasibility of labelling emotion transitions rather than emotion state, this paper aims to initiate a discussion of practical issues and open questions triggered as we proceed to classify and analyze an unusual dataset with triangulated labels: (1) resolving classification details like label resolution differences (continuous vs discrete) and data instance granularity (size of data windows); and (2) considering what it means to have confidence in the consistency or accuracy between different types (passes) of data labelling.

CCS CONCEPTS

• **Human-centered computing** → **User models**; **Empirical studies in HCI**; **Laboratory experiments**.

KEYWORDS

affect classification, emotion self-report, multimodal emotion elicitation

1 INTRODUCTION

A system that recognizes human emotion requires training data that reflects authentic spontaneous emotion, as well as reliable labels which accurately reflect actual emotion. However, ensuring the authenticity of emotion and labeling it precisely bring challenges at all phases. In a quandary well-known to emotion researchers, capturing the emotions that we express in the course of our regular lives is intrusive or impossible "in the wild"; conversely, when we elicit emotion in a lab setting, it is hard to guarantee authenticity. In either case, labelling emotion data necessitates assigning ground truths to subjective and individual experiences as if they are objective quantities.

Researchers have distilled data that captures an emotional moment into a single label by assigning an emotion word [8, 11]; or by locating the moment on a continuous dimensional space such as Russell's valence-arousal circumplex model [2, 9, 12]. While labels of this kind offer a direct path for emotion classification, it is not straightforward to select appropriate class labels. For example, multiple, even conflicting, emotions may arise from a single event; the language for communicating emotional experiences can feel woefully insufficient; identifying our feelings can require interpretation and reflection which takes time; emotions can evolve or resolve to something very different by the time we are able to verbalize the experience [3, 5, 10]. All of these suggest that a process **relying on a single label to capture a complex emotion experience from a single source minimizes richness or clarity**, relative to a reflection-based process [3]. To capture true emotional experiences and generate labels that are both accurate and rich, **we are developing a multi-pass procedure in hopes of triangulating the short-lived emotional space occupied by momentary transitional experiences**. Participants are first taken through an emotionally intense experience (for authenticity) which is physiologically and video recorded with minimal cognitive demand. In two subsequent labelling passes, participants review recordings of their original experience and annotate the timeline with two data formats: first, verbal description and word-application elicited in an interview, then a continuous emotion annotation of a 1-dimensional rating applied between the extremes of Stressed to Relaxed (a technique inspired by [6]). These labels incorporate valuable user reflection, interpretation, and interaction between mood and incited emotion, however, they also require resolving inconsistencies arising from annotating distinct data streams at different times.

We see this workshop as an opportunity to pose active questions about this approach and generate rich discussion rooted in what we have learned. The rest of this paper outlines the experiment and generated data types, then describes practical issues for creating useful data instances and finally considers the problem of resolving multiscale emotions labels for classifying a triangulated dataset.

2 BACKGROUND: METHODOLOGY AND DATA TYPES

Our open questions are based on ongoing analysis of real data collected as part of a project to detect and classify **emotion transitions**: can a sensor-based algorithm determine whether we are *becoming* more or less Stressed in a given moment? Here, we describe the data and associated formats with enough detail to ground a productive discussion. A more exhaustive description is planned for future publication.

2.1 Experiment Objectives

We designed an experiment to classify **emotion transitions – the change from one emotion to the next** that we propose may be thought of as the change in emotion over time. To elicit spontaneous emotion as it would naturally arise, participants (N=20) played an intensely disquieting animated video game (*Inside*[1]), chosen for its anxiety-provoking tension punctuated with moments of accomplishment and satisfaction, produced without depicting violence too graphically and with easy-to-learn keyboard controls.

[1]Developed by Playdead, Denmark. https://playdead.com/games/inside/

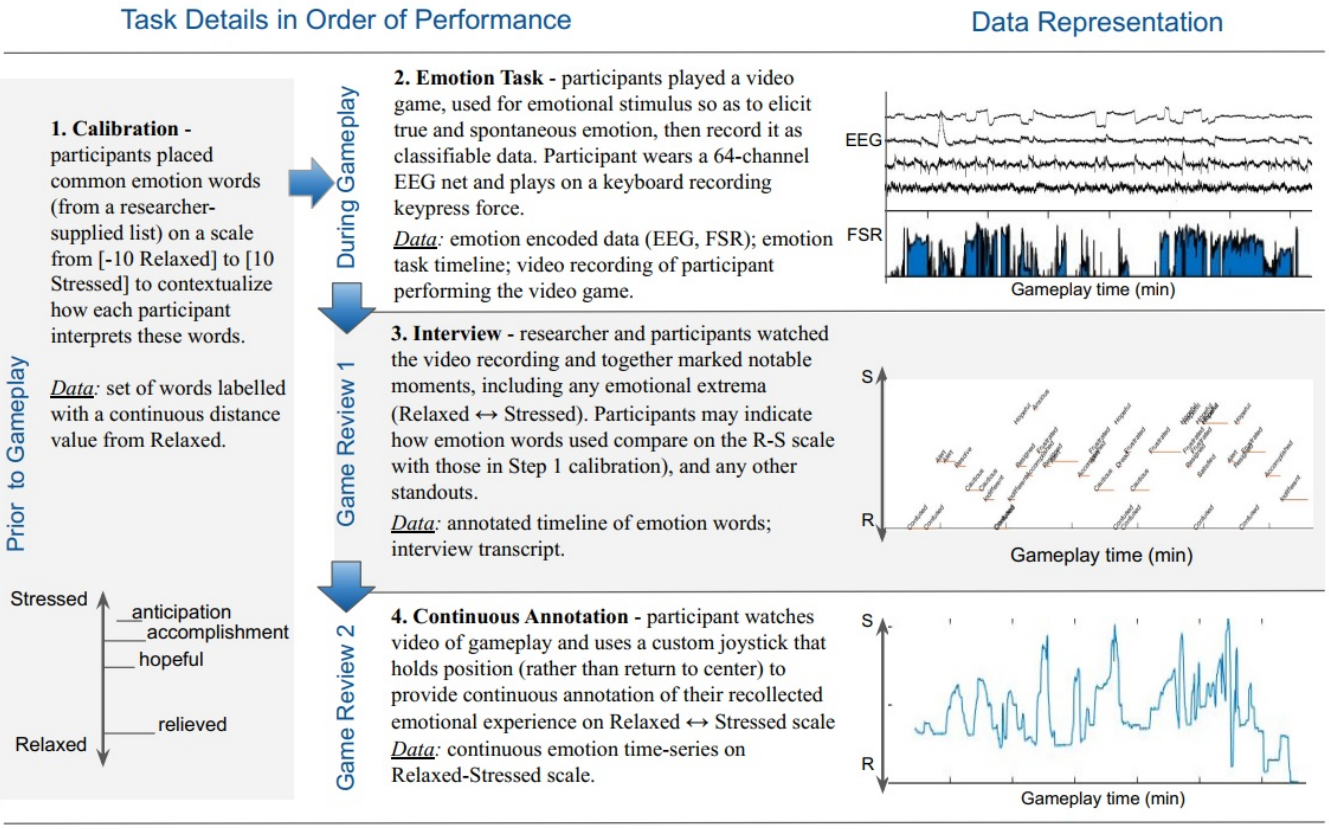

Figure 1: Participant's tasks and resulting data.

Here we outline the procedure (stages which separate the "live" experience from self-report annotation passes immediately afterward) and describe the data that is to comprise our ground truth labels: video of participants and their screen, emotion-encoded data – e.g., electroencephalography (EEG) traces or keystroke pressure – and the two types of self-report annotation described above.

2.2 Data Collection and Self-Report

We wanted participants to reconstruct how they felt on a scale between Relaxed and Stressed while engaging in a tense interactive experience. We collected 64-channel EEG data for its known ability to capture emotive properties [1]. We recorded analog pressure of keylogger activity by placing force-sensitive resistor (FSR) sensors under the standard directional movement keys (use of which is frequent and necessary to control character motion). To avoid intruding on the original emotional experience, we collected self-report emotion labeling in passes *after* the emotion task by reviewing participants' recorded facial expressions and game screen.

Figure 1 portrays the procedure's four steps. Participants (1) **calibrated** 10 commonly used, researcher-supplied emotion words along a Relaxed-Stressed axis (each emotion word is given a score based on how far along the scale (physical distance) it was placed); (2) engaged in the **emotion task** of live gameplay where emotion encoding data (EEG, FSR) was collected along with a video recording of the gameplay for review in later stages; (3) were **interviewed** and collaboratively annotated notable emotions on the gameplay timeline; and (4) **continuously annotated** their recollected emotion with 1-D joystick movements, generating a time-series of how they felt between Relaxed ↔ Stressed. The **calibration** phase (Step 1 in Figure 1) contextualizes individualized interpretations of emotion words projected onto the Relaxed ↔ Stressed axis. The remaining stages are synchronized along the original gameplay timeline such that emotion reports in the **interview** and **continuous annotation** are aligned with the **emotion task** and associated data.

In the **emotion task** (Step 2, the gameplay) where emotions are unfolding spontaneously and as a direct result of interaction with the stimulating environment, we video-recorded their screen (which depicts both the video game play and accompanying facial responses) for later review as well as select biofeedback data intended for classification purposes.

In the **interview** (Step 3), participants described to a researcher how they felt in their own words while reviewing the recording of the gameplay. The researcher paused the video and asked prompting questions where details felt important or useful. Because participants had previously undergone a word calibration phase, they had been through the exercise of considering how they considered emotions that ran from Relaxed to Stressed. We expected them to

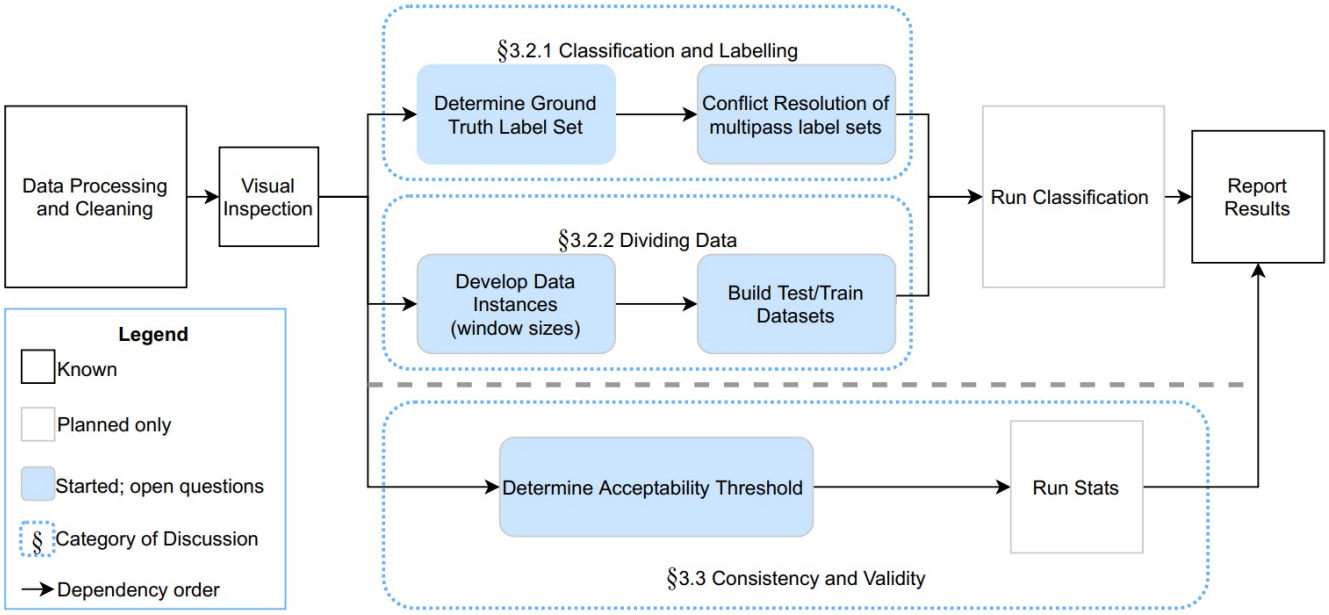

Figure 2: Analysis roadmap. For the blue boxes, identifying the best process raises intriguing questions.

be somewhat primed to use this vocabulary more frequently and more carefully than they may otherwise have done.

Finally, in **continuous annotation** (Step 4), participants used a custom joystick (holds position rather than return to center) to mark their entire emotional journey while reviewing the gameplay video yet again – this time, in its entirety and without pause. The data is a continuous time-series (256Hz) equal in duration to that of the original gameplay.

Long emotion experiments can be exhausting. To relieve participants' cognitive load, we carefully chose task order to scaffold learning on both the emotion paradigm – sorting emotions on a *Stressed* ↔ *Relaxed* scale – and the continuous annotation tool – instant and continuous joystick annotation. Neither **calibration** nor **interview** phases are time constrained, giving participants ample time to reflect and think, building up to continuous annotation (the task requiring the fastest reaction time). A learning effect is acknowledged; this is a tradeoff with prioritizing the participant experience during a challenging ~2-hr time commitment.

To assess consistency between evaluation phases, we integrated the set of calibrated words (Step 3) with the timeline of their appearance in the interview to create another annotated time-series of emotion values. With this, henceforth refer to as **Calibrated Words**, and Step 4's **Continuous Annotation**, we obtain two sets of emotion self-report aligned on the same gameplay timeline.

3 OPEN QUESTIONS

This project's goal is to assess the viability of classifying emotion transitions with a multipass labelling procedure. Here, we first outline our analysis then discuss where more consideration is needed.

3.1 General Classification Approach

We approached this analysis as we would any affect classification, by developing data instances and associated labels (Figure 2), and know that we can complete this path by running various classifiers. However, because of the multiple and sometimes conflicting data sources, we encountered novel procedural issues with respect to defining data instances and train/test sets.

Following procedures similar to related emotion classification [2, 9, 11], we used our collected data to form data instances to train a classifier that detects a participant's emotion. For transition, we project Relaxed ↔ Stressed onto a [-10, 10] scale such that: "getting Stressed" has a positive slope (changes to more Stressed) and "becoming Relaxed" has negative slope (change to less Stressed) over some time interval.

As we build these data instances, two clusters of questions arise. First, in order to classify instances, we must decide how to define labels, window sizes, and data sets. What implications will this have on interpreting accuracy measures? Second, merging multipass data labels suggests that we trust that the data is consistent and valid. What can give us the confidence that this is the case? What do we do when such a threshold is not met?

3.2 Classification Details

We have considered both discrete classifiers like Random Forest and continuous predictors like Linear Regression. For the present purpose, we focus on discrete classification for emotion state and emotion transition, as that is where we have invested the most effort as of this writing.

3.2.1 What are the implications of *discretizing* classification labels for classification…

The densest and most complete emotion annotations are generated in Continuous Annotation (final collection phase, Step 4), as it continuously marks the entire emotional task. We consider ways in which we could use this data as the ground truth.

(1) **of emotion state?**
The Relaxed ↔ Stressed scale spanning [-10, 10] can be naturally broken into at least three bins; for example, ranges for Relaxed [-10, -3 1/3); Middling (-3 1/3, 3 1/3); and Stressed (3 1/3, 10]. Presuming that data is roughly evenly distributed and that in general, Stressed follows Relaxed and vice versa, classification accuracy could be compared to chance at 33.3%.

On one hand, binning in this way makes it easier to think about the results. We can directly compare classification accuracy rates with chance, and manipulate classifier parameters for optimal results.

However, this desirable simplicity flattens much of the richness from the data, particularly in regards to how emotions move from one state to the next. By binning for even label distribution, we may overlook where emotions 'dwell' as well as interesting temporal dependencies – does Middling-Stressed pull to Stressed more often than the other way around? And how should we handle/interpret distribution changes that occur with finer bins (e.g., high instance counts of Stressed 10 and few in Stressed 4, yet these would be binned together)?

(2) **of emotion transition?**
We designed the experiment with transition in mind, trying to incorporate the temporal effects into our detection or classification system. We propose labelling based on Continuous Annotation slope, binning $\Delta emotion/\Delta t$. We found that the distribution here was not as straightforward as the slope ranges dominated near 0. To understand the impact of classifying transition *vs.* state, we tried 3 bins (to keep chance at 33.3% for ease of comparison). The quandary here is that variably-sized bins can accommodate data distribution, but do not allow for straightforward interpretation of classification results.

(3) **of some combination of state and transition?**
Classifying for state (on value) and for transition (on slope) does not account for the dynamism of the emotional experience. Perhaps feeling Stressed is more likely to pull towards more Stressed. We propose a set of combination labels of State-slope direction: e.g., Stressed+ represents the case where the data instance starts in a Stressed range value and has an overall positive slope. We hope that combining these label sets help with clarifying the experience but we worry that doing so may serve to amplify both concerns.

All of these labels are derived solely from Continuous Annotation data. The Calibrated Words data lends itself to discrete labels as it is itself a discrete and sparse dataset; so we consider the possibility of integrating both data sources to form a richer label set. After connecting the data points, we see many similarities in the overall shape of the connected Calibrated Words and the Continuous Annotation.

3.2.2 Are there interesting ways of dividing data into ...

Tantamount to considering what to label a data instance, we must consider what constitutes a data instance, training and test sets. We must consider how to divide the data for classification and subsequent interpretation of performance results.

(1) **Data windows for classification?**
An important part of considering the distribution of labels is choosing an appropriate window for creating the data instance. We chose 1s windows for being longer than the time required for human cognition (300ms after stimulus [13]), as we hope to capture the emotion expression due to an emotional event as well as the emotion prior to the event and the some of the evolution afterwards. We are considering the tradeoffs between employing overlapping windows to assess emotion evolution as a moving average *vs.* adjacent windows for accuracy comparison without overfitting the model.

However, we wonder if there are other creative ways to generate data instances that are appropriate for emotion transition labels. Since some emotions can resolve quickly while others linger, it might be interesting to select varying window sizes wherein the overlapping data will not have a great effect if using high-level statistics as features.

(2) **Training and test sets?**
We consider subject-independent classification or Leave-One-Out approaches where a classifier is trained on all data except for that of one participant and then tested on the omitted subset. Emotion expression is so individualistic that it can be used to identify an individual through this channel [2, 4, 7], raising questions about generalizable models. To make classification more meaningful, we consider other ways to partition training and test sets. Cross-validation of training is often used to inform parameter tuning and helps to determine an optimal tuning. Thus, in creating a CV performance metric, we carefully considered how to divide samples. Random division is convenient and useful for comparing classifiers but may not predict how well it would perform under real use, where only past data could inform present data. We look at other methods to construct a fuller picture.

We consider a leave-one-scene-out data partition where we set aside a test set consisting of an entire scene of the video gameplay (based on some previously determined scene demarcation procedure – e.g., set pieces, changes in game logic/puzzle completion, atmosphere, screen transitions). Training then could either use only scenes that are played prior to the test scene – a realistic but also more challenging condition in terms of guaranteeing that emotions have previously appeared in the earlier training scenes – or all other scenes regardless of their temporal presentation. Beyond these divisions, we wonder if there are other practical choices that we are overlooking.

3.3 Evaluation of Consistency

Practically, we know that many elicitation events are not conducive to simultaneous reporting; one cannot focus on playing a video

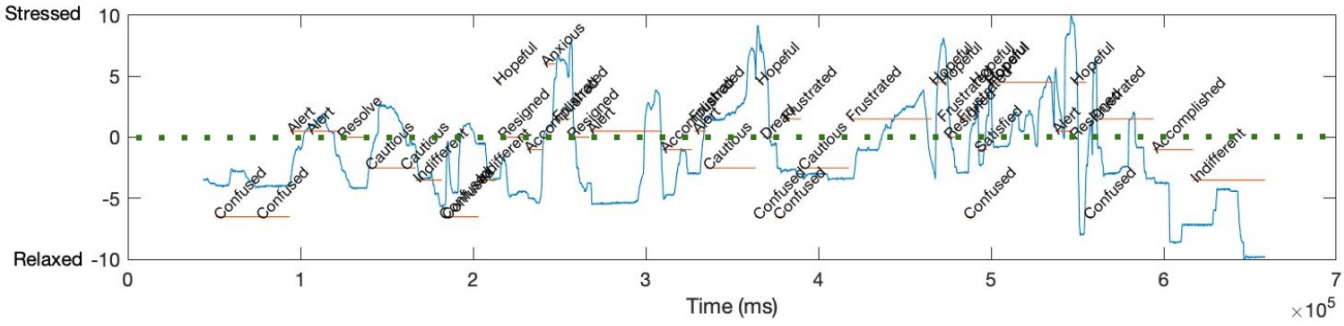

Figure 3: Continuous Annotation (blue trace) and Calibrated Words overlaid on the gameplay timeline (P02). Each word's time and Relaxed-Stressed value is aligned with a short red horizontal line and its initial letter. The y=0 axis is marked with a dotted line.

game and report unfolding and evolving emotions without negatively impacting the quality of the experience. Thus, participants are often asked to self-report on their emotions after an elicitation task is complete. When this report is performed in a single pass, we generally accept the assumption of a valid ground truth – that the participant indeed felt this way and that labels generated from these reports are consistent with the emotional expression. After all, there is no other data with which to compare. In a multipass self-report, however, we have an opportunity to quantify label validity by measuring emotion report consistency from one review to the next. Now we wonder: how closely must Continuous Annotation data and the Calibrated Words agree to provide confidence in the label? Our initial efforts compare these passes in raw data (analogous to emotion states) vs. calculated slope direction (emotion transition).

In initial visual inspection of Figure 3, which is representative of much of our data, overlaying results of the two post-game self-report passes for single representative participant's data, it appears that Calibrated Words follow a similar pattern as that of the Continuous Annotation, with most of the words landing in the same third as the Continuous Annotation value at that time point (e.g. Alert at 1 min is on the same side of the Relaxed-Stressed (or y) axis as the blue Continuous Annotation line at the same time).

To characterize the distance between these data sources, we compared the Continuous Annotation value (analogous to emotion state) with the Calibrated Word at that time point and found there to be sign agreement 68.6% of the time for data from all participants. Pearson's correlation of the same results in a correlation coefficient of 0.25 – weak correlation ($p \ll 0.001$). Then, we tried the same comparisons (sign agreement and Pearson's correlation) for Continuous Annotation slope (emotion transition) with the same time point's Calibrated Word and found 61.6% sign agreement across all participants and a moderate Pearson's correlation coefficient of 0.34 ($p \ll 0.001$). Though closer examination is necessary, we are encouraged by the result that emotion transition labels of slope may roughly approximate the commonly used state labels in terms of following people's understanding of emotion words. At minimum, this raises the questions: (1) what is the intended paradigm represented by these labels and do they reflect what people mean when they discuss their experience? And (2) what are reasonable efforts to resolve label conflicts? We currently take Continuous Annotation labels as the reference label; another approach could be to use a weighted average.

We recognize that there is more work to be done in determining an acceptability threshold for internal consistency of a multi-pass self-reporting procedure. We proposed a multi-pass post-task reporting procedure to capture experiential richness and build confidence in triangulation. But this opens up a host of questions: what might be lost when we allow for multiple post-hoc reflections of an experience to be used as ground truth labels?

4 CONCLUSIONS

We elicited an emotional experience in the lab by having participants play a tense video game, after which they engaged in self-report of their emotions, traced by both a calibrated word exercise and a continuous annotation procedure. As many emotionally rich tasks are not conducive for concurrent emotion reflection, we propose that such a multi-pass reflection exercise may be useful for triangulating the emotional experience, for example for the purpose of validating models for future real-time emotion estimation from data available intensively in real-time.

For the purposes of this workshop, we reflect on the question of ***consistency*** between review passes. These questions only arise when we consider multiple data perspectives rather than relying on one, because they introduce the possibility of conflict or complexity. We also discuss implementation decisions when using these reports as labels for classification tasks, and consider the implications for interpreting performance metrics. In particular, we are curious about the potential for using ***emotion transition*** (contextualized slope at a time point) over the more popular but temporally ill-defined emotion state (value of self-report emotion).

5 ACKNOWLEDGMENTS

This work is born from the efforts of a highly-valued team from project conception to experiment design to data collection and analysis; among others, Laura Rodgers, Hailey Mah, and Anushka Agrawal contributed through a number of evolutions of this work. We are particularly grateful to Paul Bucci for his contribution at all levels and we thank NSERC for the funding to make it possible.

REFERENCES

[1] Soraia M Alarcao and Manuel J Fonseca. 2017. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10, 3 (2017), 374–393.

[2] Kerem Altun and Karon E MacLean. 2015. Recognizing affect in human touch of a robot. *Pattern Recognition Letters* 66 (2015), 31–40.

[3] Paul H Bucci, X Laura Cang, Hailey Mah, Laura Rodgers, and Karon E MacLean. 2019. Real Emotions Don't Stand Still: Toward Ecologically Viable Representation of Affective Interaction. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Cambridge, UK, 1–7.

[4] Xi Laura Cang, Paul Bucci, Andrew Strang, Jeff Allen, Karon MacLean, and HY Sean Liu. 2015. Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 147–154.

[5] Pilar Carrera and Luis Oceja. 2007. Drawing mixed emotions: Sequential or simultaneous experiences? *Cognition and emotion* 21, 2 (2007), 422–441.

[6] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

[7] Anna Flagg and Karon MacLean. 2013. Affective touch gesture recognition for a furry zoomorphic machine. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*. 25–32.

[8] Matthew J Hertenstein, Dacher Keltner, Betsy App, Brittany A Bulleit, and Ariane R Jaskolka. 2006. Touch communicates distinct emotions. *Emotion* 6, 3 (2006), 528.

[9] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence* 30, 12 (2008), 2067–2083.

[10] Margaret McRorie and Ian Sneddon. 2007. Real emotion is dynamic and interactive. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 759–760.

[11] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[12] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[13] Rik van Dinteren, Martijn Arns, Marijtje LA Jongsma, and Roy PC Kessels. 2014. P300 development across the lifespan: a systematic review and meta-analysis. *PloS one* 9, 2 (2014), e87347.